Analyse de la valeur ajoutée d'un système de communication dans un environnement radio acoustique virtuelle (RAVE)

par

Gabriel OUAKNINE-BEAULIEU

MÉMOIRE PAR ARTICLES PRÉSENTÉ À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
COMME EXIGENCE PARTIELLE À L'OBTENTION DE LA MAÎTRISE AVEC MÉMOIRE EN GÉNIE ÉLECTRIQUE
M. Sc. A.

MONTRÉAL, LE 29 AOÛT 2025

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
UNIVERSITÉ DU QUÉBEC

PRÉSENTATION DU JURY

CETTE THÈSE A ÉTÉ ÉVALUÉE

PAR UN JURY COMPOSÉ DE:

M. Pascal Giard, directeur de mémoire
Département de génie électrique à l'École de technologie supérieure

Mme. Rachel Bouserhal, codirectrice
Département de génie électrique à l'École de technologie supérieure

M. Richard Arsenault, président du jury
Département de génie de la construction à l'École de technologie supérieure

M. Dominic Deslandes, examinateur externe
Département de génie électrique à l'École de technologie supérieure

ELLE A FAIT L'OBJET D'UNE SOUTENANCE DEVANT JURY ET PUBLIC

LE 22 AOÛT 2025

À L'ÉCOLE DE TECHNOLOGIE SUPÉRIEURE

AVANT-PROPOS

En tant qu'ancien officier mécanicien de la marine marchande, j'ai passé plusieurs années à travailler dans les salles des machines de navires commerciaux. J'y ai rencontré de nombreux marins, jeunes comme expérimentés, qui retiraient à tort leurs coquilles pour essayer de comprendre ce que leurs collègues disaient, dans un environnement assourdissant. Je suis convaincu que l'industrie maritime n'est pas la seule confrontée à ce problème. Il y a un vrai besoin d'une protection auditive qui permette de communiquer facilement avec les autres ; je l'ai vécu personnellement. Éventuellement, une protection auditive **accessible** facilitera les communications dans les milieux industriels. Ce travail est ma modeste contribution, qui vient s'emboîter dans un ensemble plus large que moi, pour la protection du public. Les petits ruisseaux deviendront de grandes rivières.

REMERCIEMENTS

Si j'ai porté la flamme de mon projet du début à la fin, je ne l'ai certainement pas fait seul. Je tiens tout d'abord à remercier mon directeur de recherche, **Pascal Giard**. Merci d'avoir pensé à moi pour ce projet, de m'avoir soutenu tout au long du parcours, tant sur le plan technique qu'administratif, et d'avoir été un appui moral jusqu'à la fin.

Je souhaite ensuite remercier ma conjointe, **Marion Lautier**. Merci de m'avoir épaulé tout au long de ce parcours : pour les bons plats que tu as préparés, pour ton soutien constant, tes encouragements inépuisables, et ta présence dans tous les moments difficiles. Bientôt, nous pourrons célébrer la fin de nos maîtrises respectives.

Je remercie **Yves Méthot**, technicien au CIRMMT. Je n'aurais jamais réalisé la collecte de donnée sans lui et son expertise technique.

Je remercie également ma codirectrice **Rachel Bouserhal** ainsi que les coauteurs de mon article, **Xinyi Zhang** et **Charles Pillet**. Merci pour votre aide précieuse dans la rédaction, la révision de l'article et l'analyse des résultats. L'expérience réalisée dans ce mémoire n'aurait pas été possible sans la participation de nombreux volontaires anonymes, que je remercie chaleureusement.

Enfin, aucun travail ne peut se concrétiser sans financement. Je remercie **Jérémie Voix**, titulaire de la chaire de recherche CRITIAS ; **EERS** ; ainsi que le **CRSNG** pour le soutien financier accordé durant ma maîtrise.

Analyse de la valeur ajoutée d'un système de communication dans un environnement radio acoustique virtuelle (RAVE)

Gabriel OUAKNINE-BEAULIEU

RÉSUMÉ

Le but de ce mémoire est d'évaluer la valeur ajoutée d'un système de communication intégrant la technologie émergente d'environnement radio-acoustique virtuel (*radio acoustical virtual environment*) (RAVE). Cette technologie, encore à l'état de concept et sans implémentation ou prototype existant, vise à améliorer les communications dans les environnements industriels. RAVE identifie les utilisateurs ciblés par un locuteur et transmet la voix électroniquement avec une impression de direction à ceux-ci.

Afin de pallier l'absence de toute réalisation, un banc d'essai émulant les principes de RAVE a été développé pour en explorer le potentiel. Il comprend trois modules : un algorithme de détection de la voix de l'utilisateur (*Own Voice Detection*) (OVD), un algorithme de détection de l'auditeur visé (*Intended Listener Detection*) (ILD) basé sur un estimateur du rayon de communication (*Communication Radius Estimator*) (CRE), et un module de spatialisation utilisant les fonctions de transfert liée à la tête (*Head-Related Transfer Functions*) (HRTFs).

Le système a été évalué avec 18 participants répartis en groupes de trois. L'analyse des questionnaires remplis par les participants et les performances des algorithmes a permis d'étudier le potentiel de RAVE.

L'OVD détectait efficacement la voix tout en libérant les mains. L'ILD, bien que fonctionnel, était peu intuitif, mais facilement améliorable grâce à un système de rétroaction. Les HRTFs étaient jugés légèrement positivement, mais leur plein potentiel n'a pas été révélé, en raison de la taille des groupes.

Ainsi, l'émulation de RAVE proposée dans ce travail a rendu possible une première évaluation : RAVE s'avère une technologie encore perfectible, mais riche en potentiel.

Mots-clés: Protecteur auditif actif, Système de communication, Traitement de signal, Environnement radio-acoustique virtuel

Analysis of the Added Value of a Communication System in a Radio Acoustical Virtual Environment (RAVE)

Gabriel OUAKNINE-BEAULIEU

ABSTRACT

The purpose of this thesis is to evaluate the added value of a communication system integrating the emerging technology of Radio Acoustical Virtual Environment (RAVE). This technology, still at the conceptual stage and without any existing implementation or prototype, aims to improve communications in industrial environments. RAVE identifies the users targeted by a speaker and electronically transmits the voice to them with a directional sound impression.

"In order to compensate for the absence of any implementation, a testbed emulating the principles of RAVE was developed to explore its potential."

In order to compensate for the absence of any implementation, a test bench emulating the principles of RAVE was developed to explore its potential. It includes three modules : an Own Voice Detection (OVD) algorithm, an Intended Listener Detection (ILD) algorithm based on a Communication Radius Estimator (CRE), and a spatialization module using Head-Related Transfer Function (HRTF).

The system was evaluated with 18 participants divided into groups of three. Analysis of questionnaires completed by the participants and algorithm performance allowed the assessment of RAVE's potential.

The OVD effectively detected the voice while freeing the hands. The ILD, although functional, was not very intuitive but can be easily improved with a feedback system. The HRTF were slightly positively received, but their full potential was not revealed due to the small group size.

Thus, the RAVE emulation proposed in this work has made a first evaluation possible : RAVE proves to be a technology that is still in need of improvement, but rich in potential.

Keywords: Active Hearing Protective Device, Communication System, Signal Processing, Radio Acoustical Virtual Environment

TABLE DES MATIÈRES

Page

LISTE DES TABLEAUX

Page

LISTE DES FIGURES

Page

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

AGC	Commande automatique de gain (*Automatic Gain Controller*)
AHPD	Protecteurs auditifs actifs (*Active Hearing Protection Device*)
ANC	Contrôle actif du bruit (*Active Noise Control*)
BWE	Extension de bande-passante (*Bandwidth Extansion*)
CIRMMT	Centre for Interdisciplinary Research in Music Media and Technology
CNN	Réseau de neurones convolutifs (*Convolutional Neural Network*)
CR	Rayon de communication (*Communication Radius*)
CRE	Estimateur du rayon de communication (*Communication Radius Estimator*)
DFT	Transformation de Fourier discrète (*Discrete Fourier Transform*)
DNN	Réseau de neurones profond (*Deep Neural Network*)
DSP	Traitement numérique du signal (*Digital Signal Processing*)
FFT	Transformation de Fourier rapide (*Fast Fourier Transform*)
FIR	Filtre à réponse impulsionnelle finie (*Finite Impulse Response*)
FPR	Taux de faux positifs (*False Positive Rate*)
HMM	Modèle de Markov caché (*Hidden Markov Model*)
HPD	Protecteur auditif (*Hearing Protection Device*)
HRTF	Fonction de transfert liée à la tête (*Head-Related Transfer Function*)
IEM	Microphone intra-auriculaire (*In-Ear Microphone*)
IFFT	Transformation de Fourier rapide inverse (*Inverse Fast Fourier Transform*)

IID	Différence interaurale d'intensité (*Interaural Intensity Difference*)
IIR	Filtre à réponse impulsionnelle infinie (*Infinite Impulse Response*)
ILD	Détection de l'auditeur visé (*Intended Listener Detection*)
ITD	Différence interaurale de temps (*Interaural Time Difference*)
LMS	Moindre carre (*Least Mean Square*)
LPC	Coefficient prédictif linéaire (*Linear Predictive Coefficient*)
MFCC	Coefficients cepstraux en fréquences de Mel (*Mel-Frequency Cepstral Coefficients*)
MMR	MultiMedia Room
NRR	Cote de réduction du bruit (*Noise Reduction Rating*)
NSSC	Centroïdes spectraux de sous-bandes normalisés (*Normalized Spectral Subband Centroids*)
OEM	Microphone extra-auriculaire (*Outer-Ear Microphone*)
OVD	Détection de la voix de l'utilisateur (*Own Voice Detection*)
PTT	Peser pour parler (*Push-To-Talk*)
RAVE	Environnement radio-acoustique virtuel (*Radio Acoustical Virtual Environment*)
SNR	Rapport voix-bruit (*Speech-to-Noise Ratio*)
SPL	Niveau de pression acoustique (*Sound Pressure Level*)
TL-HRTF	Fonction de transfert relie aux têtes de l'auditeur et du locuteur (*Talker-Listener Head Related Transfer Function*)
TPR	Taux de vrais positifs (*True Positive Rate*)

VAD	Détecteur d'activité vocale (*Voice Activity Detector*)
WIDs	Perturbations induites par le porteur (*Wearer Induced Disturbances*)
WIDsD	Détecteur de perturbations induites par le porteur (*Wearer Induced Disturbances Detector*)
ZCR	Taux de passage par zéro (*Zero-Crossing Rate*)

INTRODUCTION

La recherche présentée dans ce mémoire de maîtrise a été réalisée entre janvier 2023 et juillet 2025 au sein de la «Chaire de recherche industrielle en technologies intra-auriculaires ÉTS-EERS (CRITIAS)». L'objectif principal de cette recherche était de mesurer la valeur ajoutée d'un système de communication reposant sur la technologie émergente d'environnement radio-acoustique virtuel (*radio acoustical virtual environment*) (RAVE). Pour atteindre cet objectif, un banc d'essai devait être conçu émulant les fonctionnalités de RAVE, testé avec des participants, et les résultats du test devaient être analysés. Ce chapitre est organisé de la manière suivante : la section 0.1 décrit le contexte du problème, la section 0.2 présente les objectifs, la section 0.4 définit le cadre de la recherche, la section 0.5 détaille la méthodologie utilisée, et la section 0.6 décrit la structure du reste de ce mémoire.

0.1 Définition du problème et contexte

Pour prévenir les effets nocifs du bruit en milieu industriel, le port de protecteur auditif (*Hearing Protection Device*) (HPD) est nécessaire. Cependant, les niveaux de bruit élevés et le port de HPD rendent la transmission de la parole difficile. Grâce à la miniaturisation de l'électronique, plusieurs fonctionnalités ont pu être intégrées dans les protecteurs auditifs actifs (*Active Hearing Protection Devices*) (AHPDs). Bien que de nombreuses recherches aient porté sur l'amélioration de la protection contre les effets nocifs du bruit dans les milieux industriels, peu d'études se sont intéressées sur l'amélioration électronique de la transmission de la parole. Une technologie émergente encore à l'état de concept, nommée RAVE, tente de relever ce défi. Son objectif est d'améliorer la communication dans les environnements industriels. RAVE se distingue des autres technologies par ses fonctionnalités destinées à la fois aux locuteurs et aux auditeurs. D'un côté, RAVE vise à permettre aux locuteurs de s'adresser uniquement aux auditeurs ciblés. La sélection est basée sur l'effort déployé par les locuteurs pour produire leur voix, appelé effort vocal, tout en tenant compte du niveau de bruit. Avec le même niveau de bruit, un effort vocal

plus important implique que des auditeurs plus éloignés seront ciblés. Avec le même effort vocal, un niveau de bruit moins important implique que des auditeurs plus éloignés seront ciblés. De l'autre côté, RAVE cherche à améliorer la voix reçue par les auditeurs en leur donnant une impression de direction. Plusieurs travaux ont été réalisés sur des composantes individuelles de RAVE, telles que des algorithmes de détection d'activité vocale, de détection des perturbations induites par le porteur, et d'amélioration de la parole. Cependant, ces travaux n'ont pas encore été intégrés dans une solution complète ayant fait l'objet de tests.

0.2 Objectif de la recherche

Cette recherche poursuit un objectif principal : mesurer la valeur ajoutée de la technologie émergente RAVE. Pour atteindre l'objectif principal, trois objectifs secondaires ont été identifiés.

Le premier consiste à concevoir un banc d'essai et y inclure les différents algorithmes de RAVE, spécifiquement adaptés à une exécution en temps réel. Ce banc doit être suffisamment abouti pour offrir à l'utilisateur l'illusion que les principales fonctionnalités de RAVE sont opérationnelles en temps réel. Ces fonctionnalités concernent notamment la sélection des auditeurs en fonction de l'effort vocal, ainsi que la spatialisation de la voix, visant à fournir une impression de direction. Par ailleurs, le dispositif doit permettre l'implication d'au moins trois participants, de manière à évaluer la capacité de distinction de RAVE. Afin de garantir une expérience d'utilisation satisfaisante, le banc d'essai doit également être conçu pour fonctionner en mode sans fil, autorisant la mobilité des participants. Pour préserver la naturalité des interactions, une latence minimale est requise.

Le deuxième objectif secondaire vise à tester le banc d'essai avec des participants dans des scénarios et constituer une base de données destinée à CRITIAS. Les scénarios doivent permettre d'évaluer les fonctionnalités de RAVE, en mettant particulièrement l'accent sur le contrôle de l'identification des auditeurs ciblés ainsi que sur la perception de la directionnalité du son. Les

scénarios doivent être réalisés dans des environnements sonores calmes et bruyants, afin d'analyser l'influence des conditions sonores ambiantes. Pendant les scénarios, les participants doivent accomplir une tâche induisant une charge cognitive, de manière à observer le comportement associé à une situation d'effort mental. Certains scénarios doivent donner une condition comparative avec un système peser pour parler (*Push-To-Talk*) (PTT), afin de fournir aux participants un point de référence. Par ailleurs, les scénarios doivent positionner les participants à différentes distances, de façon à examiner la robustesse des performances de RAVE en fonction de l'éloignement des locuteurs. La base de données créée suite à l'essai par les participants permettra d'évaluer les performances des différents algorithmes de RAVE et possiblement d'autres systèmes de communication intégrés à des AHPDs. Elle sera conservée dans la banque de données nommée CRITIAS :DB et pourra être utilisée dans le cadre de projets futurs pour tester et améliorer les algorithmes de RAVE.

Enfin, le troisième objectif secondaire est de formuler des pistes d'amélioration pour les itérations futures de RAVE, et plus largement pour les systèmes de communication intégrés aux AHPDs. Cet objectif vient déterminer les points forts et les points faibles de RAVE et ultimement répondre à l'objectif principal.

0.3 Hypothèses et postulats

Bien qu'imparfait, le banc d'essai devrait fournir une bonne indication des performances de RAVE en reproduisant certaines de ses fonctionnalités. Ainsi, les participants qui utilisent le banc d'essai devraient pouvoir se faire une idée de son comportement. En recueillant l'appréciation de ces participants, des tendances devraient se dégager quant à la valeur ajoutée de RAVE. Leurs évaluations subjectives devraient refléter leur appréciation des différentes fonctionnalités ainsi que les points à améliorer. Pour mettre l'évaluation en perspective, une comparaison avec un système PTT devrait permettre de situer l'appréciation des participants. En optimisant les fonctionnalités grâce aux enregistrements réalisés lors de la collecte de données, des seuils de

performance peuvent être identifiés. Ces seuils permettent d'évaluer RAVE d'un point de vue plus objectif. En combinant les évaluations subjectives et objectives, un portrait plus complet de la valeur ajoutée de RAVE peut être établi.

0.4 Cadre de la recherche

Ce travail de recherche s'inscrit dans un projet de plus grande envergure actuellement en développement au sein de CRITIAS. À ce titre, il s'appuie sur des travaux existants, dans le but d'en prolonger la portée et de les intégrer de manière cohérente. La recherche se concentre donc sur l'intégration de ces travaux et leur adaptation à un fonctionnement en temps réel. L'objectif n'est pas de concevoir de nouveaux algorithmes plus performants, mais plutôt d'adapter des solutions existantes pour un fonctionnement en temps réel, tout en en facilitant l'analyse. De plus, les algorithmes utilisés pendant la collecte de donnée n'ont pas été optimisés pour chaque participant avant la collecte de données. Les paramètres des algorithmes étaient génériques pour tous les participants.

0.5 Méthodologie

Ce travail de recherche s'est accompli en trois grandes étapes. Premièrement, un banc d'essai a été créé, complétant ainsi le premier sous-objectif. Ensuite, ce banc d'essai fut essayé avec des participants dans une collecte de donnée, permettant de répondre au deuxième sous-objectif, soit la constitution d'une base de données. Finalement, une analyse de la collecte de données a permis de fournir des pistes d'amélioration pour les futures itérations de RAVE, complétant le troisième objectif secondaire de fournir les points forts et les points faibles de la technologie.

0.5.1 Banc d'essai

L'implémentation d'un banc d'essai pour RAVE a pour but de passer d'une technologie encore à l'état de concept et sans implémentation ou prototype existant, à une technologie qui peut être essayée. Aucune implémentation fonctionnelle des différents algorithmes ou module n'était disponible. Donc l'implémentation du banc d'essai inclut l'implémentation de ces composantes permettant les différentes fonctionnalités de RAVE. L'implémentation complète de RAVE nécessite plusieurs fonctionnalités. Pour détecter automatiquement lorsqu'un utilisateur parle, un module de détection de la voix de l'utilisateur (*Own Voice Detection*) (OVD) est requis. Réciproquement, pour identifier l'auditeur visé, un module de détection de l'auditeur visé (*Intended Listener Detection*) (ILD) est nécessaire. Puisque la voix est enregistrée dans un environnement bruyant, un filtre de bruit est indispensable. La technologie RAVE est conçue pour être utilisée avec des microphones intra-auriculaire (*In-Ear Microphones*) (IEMs). Ces microphones nécessitent un algorithme d'extension de bande-passante (*Bandwidth Extansion*) (BWE) afin d'assurer une qualité sonore satisfaisante. Pour fournir une impression de direction au son, un module de spatialisation est requis. Enfin, l'intensité du son diffusé par les écouteurs de RAVE doit être automatiquement ajustée : réduite lorsqu'elle est trop élevée et augmentée lorsqu'elle est trop faible. Ce comportement est contrôlé par une commande automatique de gain (*Automatic Gain Controller*) (AGC). Les connexions entre ces différents modules sont résumées dans la figure 0.1.

Pour résumer, les modules nécessaires au fonctionnement de RAVE sont les suivants :

- un module d'OVD ;
- un module d'ILD ;
- un filtre de bruit ;
- un algorithme de BWE ;
- filtre de spatialisation ;

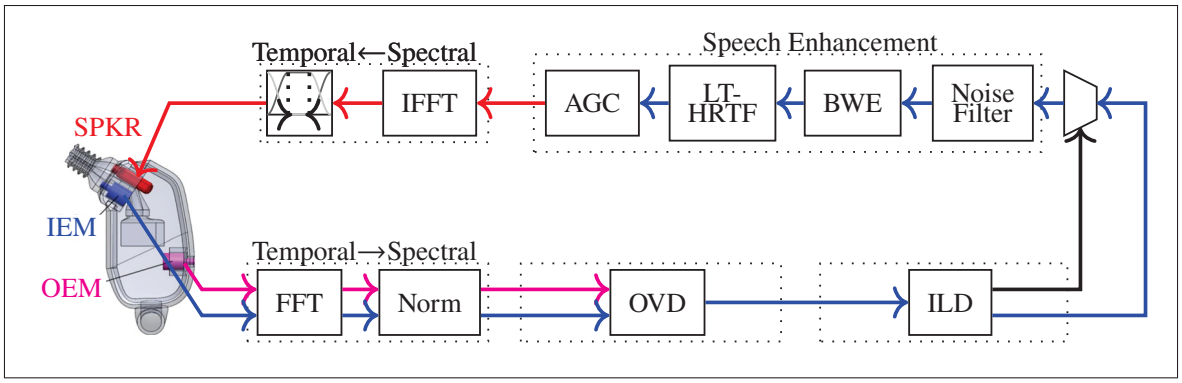

Figure 0.1 Connections des différents modules nécéssaire à RAVE

- un module d'AGC.

L'utilité et le fonctionnement de ces différentes composantes sont discutés dans la section 2.3.

Pour simuler RAVE dans un environnement industriel, le banc d'essai doit également pouvoir diffuser du bruit d'usine. Afin de faciliter son implémentation, certaines astuces ont été utilisées pour simplifier le filtre de bruit et l'algorithme de BWE. Le banc d'essai exploite l'audio d'un microphone extra-auriculaire (*Outer-Ear Microphone*) (OEM) et diffuse le bruit directement dans les oreilles des participants, plutôt que dans l'ensemble de la pièce. Les simplifications apportées ainsi que les détails de l'implémentation sont présentés plus en détail dans la sous-section 2.4.4. Ce banc d'essai, conçu à cette étape, rend possible la phase de collecte de données pour l'étape suivante.

0.5.2 Collecte de données

La collecte de données a été réalisée auprès de 18 participants, conformément à un protocole approuvé par le comité d'éthique de l'ÉTS (H20231110). Avant le début des expérimentations, l'éligibilité des participants a été vérifiée, et ceux-ci ont lu et signé le formulaire d'information et de consentement présenté en appendice. Ceux-ci ont été répartis en groupes de trois. Chaque groupe a testé le banc d'essai dans quatre configurations expérimentales distinctes :

- le banc d'essai de RAVE dans un environnement bruyant ;
- le banc d'essai de RAVE dans un environnement calme ;
- un banc d'essai d'un système PTT dans un environnement bruyant ;
- un banc d'essai d'un système PTT dans un environnement calme.

À l'issue de chaque scénario, les participants ont rempli un questionnaire évaluant la facilité d'utilisation, l'efficacité perçue et le réalisme du système testé.

Pendant la collecte de données, chaque participant disposait chacun d'un jeu de blocs ainsi que d'un ensemble de 24 instructions à réaliser. Certains blocs nécessaires à l'exécution des instructions étaient volontairement absents, une instruction sur 2, mais disponibles auprès des autres membres du groupe, favorisant ainsi les interactions. Les participants devaient s'échanger leurs blocs pour pouvoir compléter leurs constructions. En guise de remerciement, chaque participant a reçu une compensation, soit sous la forme d'un montant de 20 $, soit d'une paire de bouchons de protection auditive conçus pour les musiciens. La collecte de donnée est décrite en détail dans la section 2.4.

Tous les enregistrements issus de la collecte de données ont été sauvegardés dans la banque de données CRITIAS :DB, conformément au protocole approuvé par le comité d'éthique de l'ÉTS (H20231110). Ces enregistrements ont été utilisés lors de l'étape d'analyse de la collecte de données et pourront également servir dans le cadre de projets futurs.

0.5.3 Analyse de la collecte de donnée

L'analyse de la collecte de données permet d'identifier des pistes d'amélioration pour les futures itérations de RAVE. Elle s'appuie à la fois sur des données subjectives — issues des questionnaires et des commentaires des participants — et sur des données objectives, fondées sur les performances des différents algorithmes utilisés. L'analyse subjective offre un aperçu de l'expérience utilisateur. Elle permet d'évaluer si le concept de RAVE est perçu comme

convivial, intuitif et pratique. L'analyse objective, quant à elle, repose sur des indicateurs de performance tels que les taux de vrais positifs (*True Positive Rate*) (TPR) et les taux de faux positifs (*False Positive Rate*) (FPR), afin d'estimer la précision des algorithmes. Dans le cadre de l'identification des locuteurs et des auditeurs, l'objectif est de maximiser les détections correctes tout en réduisant les erreurs de classification. La méthode d'analyse est décrite plus en détail dans la sous-section 2.4.5.

L'analyse de la collecte de données a permis de compléter le dernier sous-objectif : identifier les points forts et les points faibles de RAVE. À la suite de cette analyse, la valeur ajoutée de RAVE a pu être mesurée sur la base de l'ensemble des étapes réalisées dans le cadre de ce projet.

0.6 Organisation du mémoire et contribution scientifique

Le travail est divisé en trois sections. Le chapitre 1 passe en revue l'ensemble des informations nécessaires à la compréhension de ce travail de recherche. Il y est notamment discuté de la nécessité des systèmes de communication intégrés aux AHPDs. Ensuite, ce chapitre survole différents travaux portant sur des systèmes de communication intégrés aux AHPDs. Le chapitre 2 correspond à un article soumis au *Journal of Acoustical Society of America*. Cet article traite de l'implémentation de RAVE, de l'implémentation du banc d'essai, de la collecte de données, et de l'analyse des résultats. Enfin, le dernier chapitre, chapitre 2.9, fait une synthèse du travail de recherche et propose des recommandations pour des travaux futurs.

CHAPITRE 1

REVUE DE LITTÉRATURE

Le présent chapitre passe en revue les connaissances essentielles à la compréhension de ce mémoire. La section 1.1 introduit les notions fondamentales liées au comportement du son, de la voix et de l'écoute, afin de poser les bases nécessaires à l'étude. La section 1.2 propose une revue de travaux antérieurs soulignant la nécessité d'un AHPD doté d'un système de communication intégré. Cette section permet de mieux comprendre l'émergence de ce besoin dans les milieux bruyants. La section 1.3 présente différentes méthodes d'évaluation associées à l'utilisation des AHPDs, en mettant en évidence les critères pertinents pour tester leur performance. La section 1.4 met en lumière les éléments clés identifiés dans d'autres systèmes de protection auditive intégrant un dispositif de communication, en soulignant l'importance de certains algorithmes spécifiques. Enfin, la section 1.5 et la section 1.6 décrivent les principes de fonctionnement de différents algorithmes et modèles pouvant contribuer à la conception d'un OVD et d'un ILD.

1.1 Le son et la voix

1.1.1 Le son

Le son est une onde de pression qui se propage dans l'air à une vitesse d'environ $344\,m/s$. Dans le Système international d'unités, la pression est exprimée en pascals (Pa). L'intensité du son la plus faible pouvant être entendue par l'humain est de l'ordre de $20\,\mu Pa$ (Driscoll, 2022). L'extrême faiblesse de cette valeur, comparée aux niveaux couramment rencontrés, justifie l'utilisation d'une échelle logarithmique. Par conséquent le niveau de pression acoustique (*Sound Pressure Level*) (SPL) d'une pression p est généralement exprimé en décibel comme exprimé par l'équation 1.1 (Driscoll, 2022).

$$SPL\,dB = 20\log_{10}(\frac{p}{20\,\mu Pa})dB \tag{1.1}$$

Les fréquences des ondes sonores audibles sont comprises entre 20 Hz et 20 kHz (Driscoll, 2022). Cependant, toutes les fréquences ne sont pas perçues avec la même intensité subjective, ce qui est lié à la notion de sonie (Driscoll, 2022). Similairement, les différentes fréquences n'ont pas le même potentiel nocif pour la santé auditive (Murphy, Kardous & Brueck, 2022). Afin de tenir compte de cet effet, plusieurs systèmes de pondération fréquentielle ont été développés. La pondération A, normalisée par la norme IEC 61672-1 :2013, est la plus couramment utilisée dans le domaine de la protection auditive. Elle est notamment adoptée par la réglementation québécoise. La conversion d'un SPL en dB vers un SPL en dBA, pour une fréquence donnée f, peut être effectuée à l'aide de l'équation 1.2, l'équation 1.3, et l'équation 1.4 (Acoustical Society of America, 2001).

$$R_A(f) = \frac{12194^2 f^4}{(f^2 + 20.6^2)\sqrt{(f^2 + 107.7^2)(f^2 + 737.9^2)}(f^2 + 12194^2)} \tag{1.2}$$

$$A(f) = 20\log_{10}(R_A(f)) - 20\log_{10}(R_A(1000)) \tag{1.3}$$

$$SPL(f)\,dBA = SPL\,dB + A(f) \tag{1.4}$$

La pondération C est également mentionnée dans la réglementation québécoise. Elle se calcule de manière similaire à la pondération A, à l'exception du facteur R (voir équation 1.2), qui diffère. Ce facteur est déterminé selon l'équation 1.5 (Acoustical Society of America, 2001).

$$R_C(f) = \frac{12194^2 f^4}{(f^2 + 20.6^2)(f^2 + 12194^2)} \tag{1.5}$$

Comme plusieurs autres types d'ondes, le SPL d'une onde sonore dépend de la distance à la source et suit la loi de l'inverse du carré. Ainsi, le SPL en dB mesuré à une distance r_1 d'une source sonore sera différent de celui mesuré à une autre distance r_2. Cette relation est exprimée par l'équation 1.6 (Driscoll, 2022).

$$SPL_2 = SPL_1 + 20\log_{10}(r_1/r_2) \tag{1.6}$$

Un son tonal correspond à un son constitué d'une seule fréquence. En pratique, la plupart des sons sont composés d'une plage de fréquences et sont généralement considérés comme des signaux à large bande. Ils peuvent être décomposés en une superposition de sons tonals à l'aide d'une transformation de Fourier discrète (*Discrete Fourier Transform*) (DFT). En pratique, une DFT est généralement implémentée avec l'algorithme de transformation de Fourier rapide (*Fast Fourier Transform*) (FFT). Le SPL en dB équivalent résultant de plusieurs ondes sonores tonal peut être calculé à l'aide de l'équation 1.7 (Driscoll, 2022).

$$SPL_{eq} = 10\log_{10}(\sum_{i=0}^{n} 10^{SPL(f_i)/10}) \tag{1.7}$$

Plusieurs ordres de grandeur existent pour donner une idée sur l'intensité de différent SPL tel que présenté dans le tableau 1.1.

Tableau 1.1 Exemple de différents SPL provenant de différentes sources de bruit (Driscoll, 2022)

Source du bruit	**SPL (dB)**
Fusil de chasse (Pointe)	158-172
Feux d'artifice (Pointe)	142-158
Cloueur pneumatique (Pointe)	116-136
Tronçonneuse	100-120
Match de football	94-114
Banc de scie	90-108
Tondeuse	74-110
Aspirateur	65-89
Froissement de papier	40-60
Écoulement de ruisseau	40-50
Chuchotement	15-40

Le son peut être composé d'un signal utile, tel que la voix, et d'un signal nuisible, appelé bruit. Le rapport entre ces deux composantes est désigné par le rapport voix-bruit (*Speech-to-Noise Ratio*) (SNR). Il est calculé avec la soustraction de l'amplitude A du signal et du bruit1.8 (Chen, Benesty, Huang & Diethorn, 2008). La compréhension de la parole est fortement influencée par le SNR. Cette capacité de compréhension est quantifiée par une mesure appelée intelligibilité.

$$SNR_{dB} = A_{signal,dB} - A_{bruit,dB} \tag{1.8}$$

1.1.2 La voix

Pour la modéliser, la voix est généralement considérée comme un signal complexe à large bande, dont les caractéristiques varient en fonction du locuteur, du contenu du message et du contexte. Le spectre vocal s'étend approximativement de 200 Hz à 8000 Hz, avec une zone

particulièrement critique entre 600 Hz et 4000 Hz, essentielle à l'intelligibilité (Casali & Tufts, 2022).

La voix est souvent modélisée comme un signal quasi stationnaire : c'est-à-dire qu'à l'échelle de courtes fenêtres temporelles, elle peut être considérée comme statique. Dans une fenêtre suffisamment petite, l'enveloppe spectrale est supposée fixe. La voix présente également une modulation en amplitude, généralement centrée autour de 3 Hz. L'intelligibilité reste bonne tant que ces modulations ne sont pas altérées dans la bande de 0.4 Hz à 20 Hz (Houtgast, Steeneken & Plomp, 1980). Il a été démontré que le filtrage des modulations inférieures à 12 Hz compromet significativement l'intelligibilité (Elliott & Theunissen, 2009).

Le spectre de la parole est constitué d'une fréquence fondamentale (ou pitch) et de ses harmoniques. Certaines fréquences, amplifiées par les résonances du conduit vocal, sont appelées formants (O'Shaughnessy, 2008). La hauteur de la voix dépend de la fréquence fondamentale, tandis que les formants dépendent des phonèmes produits. L'ensemble des phonèmes forme les mots, et donc la parole.

La production vocale est soumise à certaines limitations physiologiques. Par exemple, il est difficile de maintenir un cri sur une longue durée. L'effort vocal représente l'énergie déployée pour produire la voix et varie en fonction de la distance entre le locuteur et l'auditeur. Lorsque l'effort vocal augmente, non seulement le SPL augmente, mais les fréquences des formants et des harmoniques tendent également à s'élever (Bou Serhal, Falk & Voix, 2013).

1.1.3 L'écoute

La voix implique non seulement un locuteur, mais aussi un auditeur qui écoute. L'écoute est un processus complexe qui mobilise plusieurs capacités : détecter, différencier, reconnaître, localiser les sons, et reconnaitre la voix. Elle joue un rôle essentiel dans de nombreux environnements professionnels, où elle est déterminante pour la sécurité et la communication (Casali & Tufts, 2022).

Pour être détecté, un son doit avoir une intensité suffisante. Les sons dont l'intensité est trop faible ne sont pas perçus. Cette caractéristique est souvent utilisée pour évaluer la santé auditive à l'aide de tests audiométriques. L'un des tests les plus couramment utilisés est l'audiométrie tonale, notamment selon la méthode de Hughson-Westlake. Ce test permet d'évaluer la limite inférieure d'intensité à laquelle un son devient audible à différentes fréquences (Acoustical Society of America, 2009). Les fréquences standard testées sont, dans l'ordre : 1000, 2000, 3000, 4000, 6000, 8000, et 500 Hz. Ce test ne constitue toutefois pas une évaluation exhaustive de la santé auditive, car il ne prend pas en compte la complexité de l'écoute, c'est-à-dire les autres capacités auditives (Casali & Tufts, 2022). Néanmoins, il s'agit d'un test rapide qui fournit une indication générale de l'état de l'audition.

La capacité à distinguer un son de sa répétition, comme une voix perçue à la fois de manière passive et par un canal électronique, est modélisée par le seuil d'écho. Ce seuil correspond au temps minimal, exprimé en millisecondes (ms), nécessaire pour que deux sons soient perçus comme distincts. Le port et le type de HPD, ainsi que le niveau de bruit ambiant, influencent directement le seuil d'écho (Lezzoum, Gagnon & Voix, 2016b). Par exemple, des bouchons d'oreilles insérés plus profondément augmentent ce seuil. De même, un environnement sonore plus bruyant élève également ce seuil. Ainsi, dans un environnement très bruyant, avec des bouchons profondément insérés, un délai plus long entre un son transmis passivement et électroniquement peut être nécessaire pour qu'ils soient confondus. Dans un bruit d'usine, le seuil d'écho pour la parole est d'environ 16 ms avec des bouchons à insertion peu profonde, et d'environ 68 ms avec des bouchons profondément insérés (Lezzoum *et al.*, 2016b). Dans un système de communication où la voix peut être perçue à la fois de manière passive et électronique, le seuil d'écho détermine la latence maximale admissible pour que celle-ci reste imperceptible.

Pour être reconnu, un son doit être suffisamment fort par rapport au bruit ambiant. Aussi, le bruit, même s'il est situé dans une plage de fréquences différente, peut nuire à la reconnaissance d'un son à cause du fonctionnement non linéaire de l'oreille. Pour comprendre ce phénomène, il faut analyser le fonctionnement de l'oreille plus en détail. Lorsqu'un son est perçu, il est capté par le pavillon de l'oreille, puis dirigé vers le tympan situé au fond du canal auditif

(C.Bielefeld, 2022). Les vibrations du tympan sont transmises à trois petits os de l'oreille moyenne : le marteau, l'enclume et l'étrier. Ces osselets transmettent ensuite les vibrations à l'oreille interne. L'oreille interne comprend plusieurs structures, dont la cochlée, qui joue un rôle essentiel dans la perception sonore. Le liquide contenu dans la cochlée vibre sous l'effet des ondes sonores, excitant ainsi des cellules ciliées qui transforment ces vibrations en signaux électriques. Ces signaux sont ensuite interprétés par le cerveau. La cochlée ne fonctionne pas de manière linéaire : les sons faibles sont amplifiés, tandis que les sons forts sont saturés (Ashmore *et al.*, 2010). Cette non-linéarité entraîne la distorsion cochléaire, qui peut générer de nouvelles émissions otoacoustiques, c'est-à-dire des fréquences supplémentaires dans l'oreille non présentes dans le son d'origine. Ainsi, des sons peuvent être altérés par ces distorsions, et certaines fréquences (plage A) peuvent devenir inaudibles à cause d'un bruit trop intense dans une autre plage de fréquences proche (plage B). Ce phénomène, combiné au comportement du cerveau, explique pourquoi un bruit intense peut masquer une voix, même si leurs fréquences ne sont pas exactement les mêmes.

La localisation du son correspond à la capacité à évaluer la direction et la distance d'une source sonore, qu'elle soit en mouvement ou non, ainsi que la direction dans laquelle elle se déplace. Cette capacité est souvent associée à la compréhension de son environnement, la conscience situationnelle (*spatial awareness*).

Le cerveau détermine la direction d'un son grâce à la différence d'intensité et de phase entre les deux oreilles. Les fréquences inférieures à 1500 Hz sont utilisées pour analyser la différence interaurale de temps (*Interaural Time Difference*) (ITD), tandis que les fréquences supérieures à 3000 Hz servent à mesurer la différence interaurale d'intensité (*Interaural Intensity Difference*) (IID) (Casali & Tufts, 2022). La géométrie du corps influence à la fois l'IID et l'ITD.

Pour estimer la distance d'une source sonore, le cerveau analyse trois indices principaux :

- l'intensité du son, qui diminue avec la distance (un son plus faible peut indiquer une source plus éloignée) ;
- le rapport entre le son direct et la réverbération, qui augmente avec la distance ;

- la puissance des hautes fréquences, car l'atmosphère agit comme un filtre passe-bas en atténuant ces fréquences (Smyth, 2019).

Par ailleurs, en présence de bruit ambiant, les sons sont souvent perçus comme plus proches qu'ils ne le sont réellement. Le mouvement d'une source sonore est évalué par le cerveau grâce aux variations de niveau (intensité) et de hauteur (fréquence) lorsque le son se déplace. Comme discuté dans la section 1.3, les cache-oreilles antibruit électroniques sont généralement associés à une diminution de la capacité de localisation du son. En effet, couvrir l'auricule réduit cette capacité (Casali & Tufts, 2022).

La capacité à reconnaître la voix est souvent caractérisée par son intelligibilité. Une voix intelligible est une voix qui peut être facilement comprise. De nombreuses expériences évaluent l'intelligibilité à l'aide de tests dans lesquels des mots monosyllabiques sont prononcés (Kryter, 1946; Howell & Martin, 1975; Lindeman, 1976; Abel, Alberti, Haythornthwaite & Riko, 1982; Giguère, Laroche, Vaillancourt & Soli, 2010). Dans ce type de test, le nombre de mots correctement compris par un participant détermine le niveau d'intelligibilité. L'intelligibilité semble influencer la sollicitation des capacités cognitives (Casto & Casali, 2013). Dans ce mémoire, cette sollicitation est désignée sous le terme de charge de travail.

1.2 La communication dans un environnement industriel

1.2.1 Pourquoi faut-il porter une protection auditive

En fonction du SPL auquel une personne est exposée, des effets nocifs, tels que des pertes auditives ou la surdité, peuvent survenir. Par conséquent, une protection auditive adéquate est essentielle dans les milieux bruyants, notamment dans les environnements industriels.

Selon la Loi sur la santé et la sécurité du travail, section 15, article 131, l'exposition quotidienne au bruit ne doit pas dépasser 85 dBA sur une période de 8 heures, et 140 dBC à tout moment, avec une pression acoustique de référence de 20 μPa (Ministère du Travail, de l'Emploi et de la Solidarité sociale, 2013). Pour limiter cette exposition, des protecteurs auditifs ou HPDs

peuvent être utilisés. Les HPDs sont des équipements de protection individuelle conçus pour atténuer les effets nocifs du bruit. Ils sont généralement caractérisés par leur forme, leur cote de réduction du bruit (*Noise Reduction Rating*) (NRR) ainsi que par leurs fonctionnalités spécifiques (H. Berger & Voix, 2021; Casali, 2010).

1.2.2 Type de protection auditive passive

Les HPDs peuvent se présenter sous différentes formes, telles que les coquilles (casques) ou les bouchons auriculaires. De manière générale, les bouchons sont associés à un effet d'occlusion plus prononcé que les coquilles. Ce phénomène se produit lorsque le canal auditif est obstrué, ce qui entraîne une amplification des sons transmis par voie osseuse, notamment nos propres sons physiologiques (respiration, mastication, voix, etc.).

L'ostéophonie, c'est-à-dire la perception des sons transmis par les os du crâne, se manifeste principalement dans les fréquences inférieures à 1 kHz (Carillo, Doutres & Sgard, 2020). Ainsi, en présence d'un effet d'occlusion, notre propre voix nous semble plus forte et plus grave, car elle est amplifiée dans les basses fréquences.

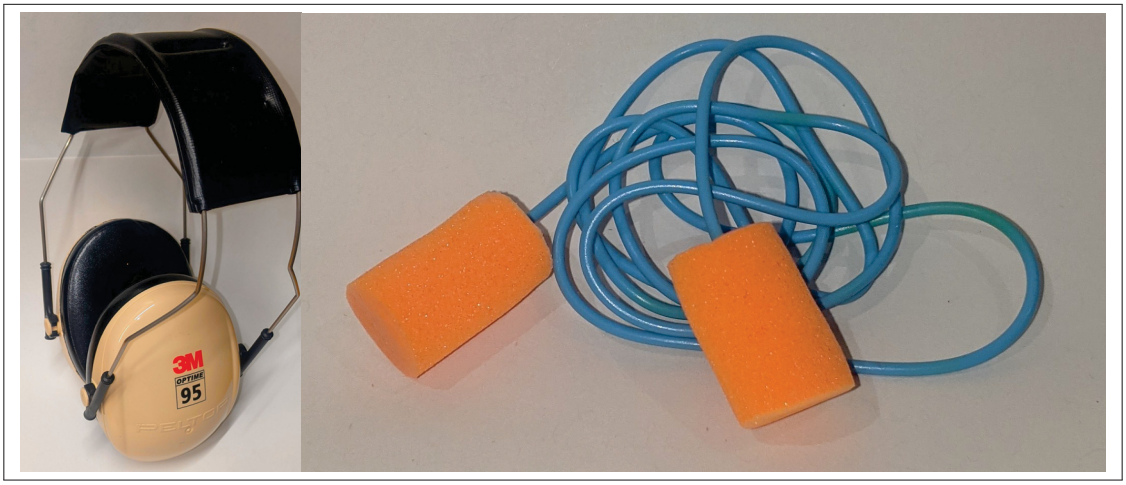

Figure 1.1 Exemple de coquilles et de bouchons

Leur NRR indique le niveau d'atténuation sonore qu'ils peuvent fournir. Selon les modèles, les bouchons d'oreilles offrent typiquement un NRR d'environ 30 dB, tandis que les coquilles présentent un NRR compris entre 15 et 25 dB.

Cependant, l'atténuation réelle offerte par les HPDs dépend fortement de comment ils sont portés, en particulier pour les bouchons, dont l'efficacité peut varier considérablement selon l'insertion dans le conduit auditif. L'efficacité est également grandement compromise lorsque les utilisateurs retirent leur protection, même pour de courtes périodes. Par exemple, un utilisateur portant des HPDs offrant 30 dB d'atténuation, mais les retirant pendant seulement 15 minutes au cours d'une journée de travail de 8 heures, bénéficie en réalité d'une protection effective réduite à 22 dB (H. Berger & Voix, 2021).

1.2.3 Problème de la protection auditive passive

Depuis longtemps, de nombreuses industries se plaignent que les HPDs réduisent l'intelligibilité de la parole. Ce phénomène est toutefois demeuré mal compris pendant plusieurs années. Retracer l'évolution de la compréhension de cet effet à travers le temps permet de mieux cerner la problématique de la communication en milieu industriel et de situer les enjeux actuels liés à l'usage des HPDs dans des environnements bruyants.

Dès 1946, Kryter a étudié la perception de la parole et de l'écoute avec le port de bouchons d'oreilles. Il cherchait à démontrer que les bouchons pouvaient constituer une solution adéquate pour la protection auditive, sans nuire de manière significative à l'intelligibilité de la parole. À l'aide d'une expérience menée sur huit hommes ayant une audition normale, il a conclu que, dans un environnement bruyant, l'intelligibilité pouvait être légèrement meilleure avec des bouchons que sans. Selon ses données, l'intelligibilité diminuait à mesure que le niveau de bruit augmentait, mais cette dégradation était comparable avec ou sans bouchons d'oreilles (Kryter, 1946).

En 1976, Howell & Martin ont étudié la parole et l'écoute dans le bruit en comparant l'effet de coquilles et de bouchons d'oreilles sur l'intelligibilité. Ils ont conclu qu'à de faibles niveaux de

bruit, les HPDs réduisent l'intelligibilité pour l'auditeur. Toutefois, lorsque le niveau de bruit ambiant se situe entre 80 dB et 95 dB, l'intelligibilité est légèrement améliorée avec les HPDs. Ils ont également émis l'hypothèse que, au-delà de ces niveaux sonores, l'intelligibilité pourrait devenir meilleure sans protection auditive. Il convient de noter que les niveaux de bruit testés dans cette étude dépassaient nettement les limites d'exposition fixées par la réglementation moderne pour une période de 8 h. Concernant la parole émise par le locuteur, les auteurs ont observé une réduction de l'intelligibilité à mesure que le niveau de bruit augmente. Ils ont proposé que les locuteurs régulent leur voix en s'appuyant sur l'ostéophonie, influencée par le phénomène d'occlusion. Ce mécanisme expliquerait les différences d'intelligibilité observées entre le port de bouchons et de coquilles chez leurs participants. Leur expérimentation, menée auprès de douze hommes ayant une audition normale, a aussi exploré l'effet de différentes distributions fréquentielles du bruit. Selon leurs résultats, l'impact sur l'intelligibilité était similaire pour les deux types de bruit testés.

La relation entre le port de HPDs et l'intelligibilité de la parole a été largement clarifiée en 1976. Cette même année, Lindeman a démontré que le port de HPDs dans un environnement bruyant réduit fortement l'intelligibilité chez les personnes présentant une perte auditive. Son étude, menée auprès de 537 participants âgés de 17 à 65 ans, comprenait des tests audiométriques ainsi que des épreuves d'intelligibilité de la parole dans différents niveaux de bruit, avec et sans HPDs. Les résultats ont permis de formaliser la relation entre la capacité auditive et l'effet des HPDs sur l'intelligibilité. Il a observé que, chez les individus à l'audition normale, le port de HPDs pouvait parfois améliorer l'intelligibilité. En revanche, chez les personnes ayant une perte auditive, le port de HPDs entraînait systématiquement une détérioration de l'intelligibilité. L'étude a également révélé que, si l'âge était bien corrélé aux pertes auditives, il ne l'était pas directement à l'intelligibilité, une fois le facteur auditif contrôlé.

En 1982, Abel *et al.* poursuivent l'étude sur la communication dans le bruit avec le port de HPD. Leur objectif était de déterminer d'autres variables importantes influençant l'intelligibilité dans le bruit. Ils se sont intéressés au profil fréquentiel des pertes auditives, à la fluidité verbale en anglais de leurs participants, au profil fréquentiel du bruit, au type d'HPD utilisé, ainsi qu'à

l'intensité de la parole. Les bouchons d'oreille offraient une atténuation plus importante entre 125 et 500 Hz, mais cela n'a pas eu d'impact significatif sur l'intelligibilité. Le port de HPD diminue toutefois fortement l'intelligibilité chez les personnes avec des pertes auditives. Ils ont également constaté que la différence de langue entraînait une perte de 10 à 20 % d'intelligibilité. De plus, le profil fréquentiel du bruit jouait un rôle important : le bruit d'une foule se révélait plus perturbant que le bruit blanc.

En 1987, Casali & Horylev ont mené une expérience sur l'intelligibilité dans le bruit. Il a fait varier le niveau d'occlusion, l'intensité du bruit, ainsi que le profil fréquentiel du bruit. Parmi les 45 participants, tous avec une audition normale, 23 étaient des hommes et 22 des femmes. Le résultat le plus marquant de cette étude concerne l'effet de l'occlusion sur l'intelligibilité. Une occlusion plus faible, apportée par des coquilles, favoriserait une meilleure compréhension dans des niveaux de bruit plus faibles, autour de 60 dBA. À l'inverse, une occlusion plus forte, induite par des bouchons, permettrait une meilleure compréhension dans des niveaux de bruit plus élevés, environ 83 dBA. Dans tous les cas, Casali & Horylev conclut qu'il vaut mieux se protéger avec des HPDs que de rencontrer des difficultés à se comprendre dans un environnement bruyant.

En 2010, Giguère *et al.* (2010) proposent un modèle pour prédire l'intelligibilité dans le bruit, avec ou sans protecteurs auditifs, en tenant compte à la fois de l'audibilité et de la distorsion. Selon ce modèle, lorsque le son est plus fort, un SNR plus faible suffit pour atteindre une meilleure intelligibilité. En revanche, pour une personne présentant une perte auditive, un SNR plus élevé est nécessaire afin d'obtenir une bonne compréhension. Ce modèle affiche d'excellentes performances, avec une erreur moyenne de 1,7 % et un écart-type de 15,5 %. Comme les recherches précédentes, il confirme que l'audition influence fortement l'intelligibilité lors du port de HPDs.

De ces différentes recherches sur la communication avec le port de HPD, plusieurs conclusions peuvent être tirées. Premièrement, plus le niveau de bruit est élevé, plus il devient difficile d'atteindre une bonne intelligibilité. Celle-ci dépend principalement du SNR. De plus, certaines

personnes ont tendance à parler moins fort lorsqu'elles portent des HPDs, ce qui réduit le SNR et, par conséquent, l'intelligibilité. Ensuite, il a été constaté que, pour les personnes avec une bonne audition, le port d'HPDs n'affecte pas significativement l'intelligibilité jusqu'à un certain seuil de bruit, seuil qui est influencé par le niveau d'occlusion. À des niveaux de bruit plus élevés, une forte occlusion favorise une meilleure compréhension. Le bruit réduit cependant toujours l'intelligibilité, mais cet effet est moindre chez les personnes sans troubles auditifs. En revanche, pour les personnes présentant une perte auditive, l'intelligibilité peut être fortement dégradée par le port de HPDs en milieu bruyant. Enfin, le type de bruit joue un rôle important : le profil fréquentiel et temporel du bruit influence l'intelligibilité. Un bruit blanc a moins d'impact négatif que le bruit d'usine, ce qui explique pourquoi de nombreuses expériences d'intelligibilité utilisent des bruits plus réalistes.

1.3 Protecteurs auditifs actifs

Avec l'avènement de la miniaturisation de l'électronique, les AHPDs ont vu le jour. Ces dispositifs peuvent offrir des fonctionnalités intéressantes adaptées à divers environnements, comme le contrôle actif du bruit (*Active Noise Control*) (ANC), la transmission de la parole, et la spatialisation du son. Dans ce mémoire, on désigne par ANC l'ensemble des systèmes visant à annuler ou réduire le bruit au moyen d'un haut-parleur et d'un microphone. De nombreuses études se sont penchées sur l'interaction entre les AHPDs et leurs usagers. Cette section synthétise certains résultats clés de recherches portant sur la perception spatiale du son ainsi que sur l'intelligibilité de la parole.

1.3.1 Intelligibilité

L'étude de Casto & Casali (2013) portait sur la charge de travail des pilotes d'hélicoptère dans l'armée. L'objectif était d'évaluer les effets combinés de différents niveaux de charge de travail, de divers systèmes de communication, ainsi que de la qualité des signaux audio. Les données ont été recueillies auprès de 20 pilotes militaires ayant des profils auditifs variés. Dans un simulateur, les pilotes devaient suivre des trajectoires prédéfinies à différentes vitesses et

altitudes. La complexité et la fréquence des consignes reçues servaient à moduler la charge de travail. La performance de pilotage était mesurée par les écarts entre les ordres donnés et les trajectoires suivies. Par ailleurs, les pilotes devaient répéter les consignes entendues, et les erreurs de répétition étaient utilisées pour évaluer l'intelligibilité de la communication. Les résultats ont montré qu'une charge de travail élevée et l'utilisation de systèmes de communication de moindre qualité étaient statistiquement associées à une baisse de la performance de pilotage et de l'intelligibilité. En revanche, les profils auditifs des pilotes n'étaient pas corrélés avec une baisse de performance ou d'intelligibilité. Les auteurs suggèrent que l'expérience accumulée par les pilotes ayant un profil auditif moins favorable pourrait compenser les effets négatifs d'une charge de travail élevée et d'un système de communication moins performant. Ils avancent toutefois que, dans des conditions de charge encore plus intense, une corrélation significative entre profil auditif et performance pourrait apparaître.

L'étude comparative de Ribera, Mozo & Murphy (2004) a évalué différents AHPDs munis de systèmes de communication, destinés aux pilotes d'hélicoptère. L'objectif était de déterminer quel système offrait la meilleure intelligibilité de la parole. Trois systèmes ont été comparés : un casque d'aviation standard (SPH-4B), un bouchon de protection muni d'un haut-parleur inséré dans le conduit auditif, et un casque de pilotage équipé d'un système d'ANC. L'étude a été réalisée auprès de 40 pilotes, dont 20 avec une audition normale et 20 présentant des pertes auditives. Après avoir administré des tests d'intelligibilité de la parole à chacun des participants pour chaque dispositif, les résultats ont permis de classer les systèmes selon leur performance. L'ordre décroissant d'intelligibilité était le suivant : (1) le bouchon de protection muni d'un haut-parleur dans le conduit auditif, (2) le casque avec ANC, puis (3) le casque d'aviation standard (SPH-4B).

En résumé, il existe un lien clair entre l'intelligibilité de la communication et la charge de travail cognitive. Une faible intelligibilité tend à accroître cette charge, rendant non seulement la communication plus exigeante, mais augmentant également l'effort mental requis. Les bouchons de protection munis d'un haut-parleur intégré dans le conduit auditif semblent constituer la solution la plus efficace en termes d'intelligibilité. Par ailleurs, le contrôle de la charge de travail

semble important lors de l'évaluation d'un système de communication, même si ce facteur demeure difficile à maîtriser expérimentalement.

1.3.2 Perception spatiale du son

L'étude militaire de Abel, Tsang & Boyne (2007) portait sur la difficulté à localiser la provenance des sons lors du port d'AHPDs. La perception spatiale du son est très importante pour les militaires lors de mission. Les participants devaient identifier la direction de bruit blanc diffusé autour d'eux. L'expérience a été réalisée dans trois conditions : sans HPD, avec des AHPDs munis d'un système de ANC, et avec des AHPDs dotés d'un système de communication. Douze participants ayant une audition normale ont été recrutés. L'étude a conclu que les AHPDs affectent la capacité à localiser les sons.

L'étude de Niermann (2015) portait sur l'utilisation de la spatialisation sonore comme aide à la réduction de la charge de travail dans les cockpits d'avion. En général, les informations destinées au pilote sont principalement transmises de façon visuelle, à travers de nombreux cadrans. De plus, les alarmes auditives traditionnelles ne fournissent aucune information directionnelle. Cette étude propose que l'intégration de sons directionnels puisse améliorer l'interface homme-machine et alléger la charge cognitive du pilote. Pour évaluer la faisabilité de cette approche, les auteurs ont testé la capacité de localisation sonore chez 23 participants, dont 10 titulaires d'une licence de pilote, tous ayant une audition normale. Les participants portaient un casque audio diffusant des sons directionnels générés à l'aide de filtres et devaient identifier leur provenance spatiale. L'écart type des erreurs de localisation était de 10°. Les résultats suggèrent que la spatialisation sonore présente un fort potentiel pour améliorer l'ergonomie des cockpits.

La perception spatiale du son peut être très importante pour les milieux industriels. Dans plusieurs milieux les alarmes de recul de véhicules sont importantes à repérer. Laroche *et al.* (2022) s'est intéressé à comment les AHPDs et les HPDs affectent la capacité de repérer la provenance d'une alarme tonale et d'une alarme a large bande. Du côté des HPDs, les coquilles combinées avec les bouchons affectaient grandement la localisation, suivi de seulement le port

des coquilles et seulement les bouchons affectait le moins. Avec des AHPDs, les coquilles munies d'ANC affectaient plus que les bouchons munis d'ANC. L'étude à aussi trouver les alarmes tonales était plus difficile à repérer que les alarmes à large bande.

En résumé, la perception spatiale du son est importante dans plusieurs domaines — militaire, aéronautique, et industriel. Dans les milieux militaire et industriel, elle est directement liée à la sécurité. Parmi les différents AHPDs testés, ce sont les bouchons qui altèrent le moins cette perception. Dans l'aviation, la spatialisation du son pourrait contribuer à réduire la charge de travail des pilotes.

1.4 Implémentation de système de communication pour les milieux industriels

Différents travaux décrivent l'intégration de systèmes de communication dans des AHPDs. Leur analyse met en lumière diverses pistes pertinentes pour guider l'implémentation de tels systèmes.

Westerlund, Dahl & Claesson (2005) décrivent un système de communication intégré dans un AHPD, combinant le port de coquilles et de bouchons. Les coquilles offrent une fonction d'ANC, tandis que les bouchons sont munis d'IEMs pour capter la voix. La voix est captée grâce à l'ostéophonie et à l'effet d'occlusion. Cet arrangement présente l'avantage de fournir un signal moins bruité, sans nécessiter l'alignement de la bouche avec un microphone, mais il comporte tout de même certains défis. La voix captée par ostéophonie n'est pas exempte de bruit, et sa signature spectrale est modifiée, comme discuté dans la section 1.1. Ce travail s'est focalisé sur plusieurs aspects : la mise en œuvre de l'ANC, l'enregistrement de la voix via un IEM, la détection de la présence de parole avec un détecteur d'activité vocale (*Voice Activity Detector*) (VAD), l'amélioration de la qualité spectrale de la voix, ainsi que la réduction du bruit résiduel. L'ANC est implémentée à l'aide d'une boucle de rétroaction analogique incluant un microphone et un haut-parleur. L'efficacité des IEMs à capter la voix a été vérifiée expérimentalement. Pour ne transmettre que les segments contenant de la parole, un VAD est nécessaire. Dans cette étude, un VAD basé sur un modèle de Markov caché (*Hidden Markov Model*) (HMM) a été utilisé. Avec ce système, 76,7 % des interactions correspondaient à des mots présents correctement détectés,

10 % à des mots présents non détectés, et 13,3 % à des mots absents, mais faussement détectés. Comme mentionné, la voix intra-auriculaire est plus forte et décalée vers les basses fréquences. Ce travail compare la voix intra-auriculaire à une version filtrée par un filtre passe-bas de la parole émise par la bouche. Donc, pour améliorer la qualité sonore, les hautes fréquences sont amplifiées à l'aide d'un filtre passe-haut. Enfin, même avec l'ANC, la présence de bouchons, et de coquilles, le signal n'est pas exempt de bruit. Pour atténuer ces bruits, un algorithme de soustraction spectrale est utilisé. Le SNR de cet arrangement est de -11 dB avec seulement les coquilles, de -4 dB avec l'ANC activé, et atteint 28 dB après application de la soustraction spectrale. Ce travail est intéressant, car il décrit bien les modules nécessaires à la transmission de la voix grâce à des IEM. Par contre la réception d'un signal vocal n'est pas abordée dans ce travail. De plus, la combinaison de bouchon et de coquille nuit à la perception spatiale du son comme discuté dans la section 1.3.

Pawełczyk, Latos, Michalczyk, Czyz & Mazur (2011) présentent un système de communication destiné à des utilisateurs regroupés, reposant exclusivement sur des bouchons d'oreilles intégrant des IEMs, des OEMs, ainsi que des haut-parleurs assurant à la fois une fonction d'ANC et de réduction du bruit. L'ANC est mise en œuvre à l'aide d'une architecture de commande anticipative, exploitant les signaux issus des IEMs et des OEMs. La réduction du bruit est quant à elle obtenue par un filtrage adaptatif fondé sur l'algorithme de moindre carre (*Least Mean Square*) (LMS), en utilisant comme référence le signal issu des OEMs, lesquels captent uniquement le bruit ambiant. Ce travail met l'accent sur les contraintes du traitement en temps réel, ce qui impose une minimisation de la complexité algorithmique. Par exemple, il est suggéré d'éviter le recours à un VAD, afin de réduire la charge de calcul. Toutefois, il est également mentionné que l'implémentation pratique du système requiert malgré tout l'usage d'un VAD, bien que son fonctionnement ne soit pas détaillé. Enfin, il est brièvement indiqué que les utilisateurs présents dans une même zone partagent un canal de communication, et que jusqu'à trois locuteurs peuvent parler simultanément à l'intérieur de cette zone, dans une configuration de type «hot spot». Ce travail est intéressant, car il propose une solution fondée uniquement sur des bouchons d'oreilles intégrant des IEMs, des OEMs, et des haut-parleurs. Il confirme

la faisabilité de l'ANC à l'aide de bouchons seuls. Cependant, ce travail présente plusieurs angles morts. Il existe une contradiction concernant la nécessité d'un VAD : bien qu'il soit recommandé de s'en passer pour des raisons de calcul, son utilisation est malgré tout considérée indispensable, sans explication sur son intégration réelle. En outre, aucune information n'est donnée quant à la manière dont le système évite d'annuler la voix d'un utilisateur ou les sons transmis dans les haut-parleurs, ce qui soulève des interrogations sur la coexistence de l'ANC et des fonctions de communication. Enfin, aucun algorithme n'est proposé pour corriger le profil spectral de la voix, altérée par les effets combinés de l'ostéophonie et de l'occlusion.

Le travail de Brammer *et al.* (2014a) s'est concentré sur le développement de coquilles de protection auditive intégrant un système d'ANC, tout en maintenant un canal de communication externe. Ce travail décrit principalement la restitution de l'audio à l'intérieur des coquilles, sans toutefois préciser l'origine de ce signal audio. Le système repose sur un microphone placé à l'extérieur des coquilles, un second à l'intérieur, ainsi qu'un haut-parleur interne. Le microphone externe alimente une boucle de commande anticipative pour l'ANC, tandis que le microphone interne est utilisé dans une boucle de rétroaction. La combinaison de ces deux boucles permet à la fois d'effectuer une annulation active du bruit et de restituer un signal audio sans compromettre la performance de l'ANC. L'objectif principal de cette étude était le développement de l'algorithme d'ANC et la démonstration d'une preuve de concept. L'intelligibilité du système a d'abord été évaluée en simulation à l'aide de MATLAB, puis testée expérimentalement avec six participants à travers un protocole d'évaluation de l'intelligibilité de la parole. Selon Brammer *et al.*, les résultats de ce test se sont révélés très prometteurs. La solution proposée est particulièrement intéressante, car elle pourrait potentiellement être adaptée à des bouchons d'oreilles intégrant des IEMs, OEMs et des haut-parleurs. Ce système décrit la partie réception d'un canal de communication dans un AHPD, tout en maintenant une capacité d'ANC, ce qui permettrait d'augmenter le NRR global du dispositif. Cependant, cette étude ne fournit pas de détails explicites concernant la transmission de l'audio par l'utilisateur, limitant ainsi l'évaluation complète de son usage dans un contexte de communication.

L'ANC apparaît comme un élément crucial d'un AHPD, car il contribue directement à l'objectif principal : la protection de l'audition. Plusieurs travaux ont étudié l'ANC, en se concentrant soit sur la fonction de transmission, soit sur la fonction de réception, mais rarement sur l'intégration des deux. L'utilisation d'un IEM semble être une piste prometteuse pour différents travaux. Le fonctionnement en temps réel est un critère crucial. Pour assurer la transmission, il est nécessaire d'intégrer un module de détection de la voix de l'utilisateur, associé à un filtre de bruit ainsi qu'à un module de correction du profil spectral (BWE).

1.5 Détection de la voix de l'utilisateur

Un système de communication intégré à un AHPD peut nécessiter un module de détection de la voix de l'utilisateur. Un OVD permet de détecter la présence d'une voix tout en la distinguant de celle d'autres utilisateurs. Les OVD sont généralement composés de deux parties : un détecteur de perturbations induites par le porteur (*Wearer Induced Disturbances Detector*) (WIDsD), qui identifie la voix de l'utilisateur en la différenciant de celles des autres, et un VAD, qui confirme la présence de la parole.

1.5.1 Détection de perturbation induite par l'utilisateur

Il existe deux grandes familles de techniques permettant de distinguer l'activité vocale de l'utilisateur : l'utilisation d'accéléromètres et celle de plusieurs microphones. Différents travaux ont proposé de positionner un accéléromètre sur la poitrine, la gorge ou le cou de l'utilisateur (Pertilä, Fagerlund, Huttunen & Myllyla, 2021). L'accéléromètre permet de détecter les vibrations générées lorsque l'utilisateur parle. Le signal d'accélération est généralement filtré à l'aide d'un filtre passe-haut afin d'atténuer les artefacts liés aux mouvements corporels. Ce signal peut ensuite être comparé à des seuils prédéfinis ou traité à l'aide d'un algorithme d'apprentissage automatique pour effectuer une classification (Pertilä *et al.*, 2021).. Cette méthode est considérée comme plus robuste que l'utilisation exclusive de microphones, bien qu'elle nécessite de distinguer les accélérations pertinentes de celles causées par le bruit, en plus d'impliquer l'ajout

d'un accéléromètre. Enfin, cette approche peut être exploitée directement dans le domaine temporel, sans nécessiter de transformation fréquentielle.

Certains travaux proposent une solution reposant sur le positionnement d'un OEM sur chaque oreille (Pertilä *et al.*, 2021; Bitzer, Bilert & Holube, 2018). La bouche est, à toute fin pratique, considérée comme située à mi-chemin entre les deux oreilles. Ainsi, la voix de l'utilisateur devrait parvenir aux deux OEMs avec une intensité similaire et en phase. La largeur de la tête est généralement suffisante pour que la voix d'autres personnes ne soit ni en phase ni de même amplitude sur les deux microphones. Les approches les plus performantes dans la littérature s'appuient sur la mesure de la cohérence entre les signaux issus des deux OEMs (Bitzer *et al.*, 2018). La cohérence quantifie le degré de corrélation entre deux signaux en fonction de la fréquence, et elle est généralement moyennée sur une plage fréquentielle donnée. Cette méthode reste toutefois sensible au niveau de bruit ambiant et à la position des interlocuteurs. Par exemple, la voix d'un autre utilisateur situé directement en face ou à l'arrière peut être captée par les deux microphones avec une intensité et une phase similaires, ce qui peut conduire à de faux positifs.

Une alternative, utilisée en audiométrie, repose sur l'utilisation conjointe d'un OEM et d'un IEM (Bonnet, Nelisse, Nogarolli & Voix, 2019). Initialement développée pour détecter les perturbations induites par le porteur (*Wearer Induced Disturbances*) (WIDs), cette méthode vise à identifier non seulement la parole, mais aussi d'autres sons produits par l'utilisateur, tels que les raclements de gorge, les clignements des yeux, etc. Elle peut donc également être exploitée pour détecter spécifiquement la parole de l'utilisateur. L'approche repose sur la mesure de la cohérence entre les signaux captés par l'OEM et l'IEM, moyennée sur une plage de fréquences définie. Développée dans le contexte du port de bouchons de protection auditive, cette méthode est conçue pour être résistante au bruit ambiant. De plus, en raison de l'étanchéité acoustique entre l'OEM et l'IEM, elle ne devrait pas être sensible à la position d'autres utilisateurs.

La fonction de cohérence nécessite une transformation des signaux dans le domaine fréquentiel. En pratique, la FFT est utilisée pour convertir un signal temporel en sa représentation spectrale. Une FFT est appliquée à un signal $\mathbf{x}(t)$ sur une fenêtre de temps allant de t_0 à t_{window}, afin d'obtenir

Tableau 1.2 Résumé des techniques de WIDsD

	Techniques	Effet du bruit	Effet de la position	Effet des mouvements
1	Accéléromètre	Résiliente au bruit	Insensible à la position des autres utilisateurs	Sensible aux artefacts de mouvement
2	2 OEM	Sensible au bruit	Sensible à la position des autres utilisateurs	Insensible aux artefacts de mouvement
3	OEM + IEM	Résiliente au bruit	Insensible à la position des autres utilisateurs	Peut détecter tout les WIDs

sa représentation dans le domaine fréquentiel, notée $\mathbf{X}(t)$, comme illustré dans l'équation 1.9.

$$\mathbf{X}(f) = \mathbf{FFT}(\mathbf{x}(t))|t \in [t_0, t_{\text{window}}] \tag{1.9}$$

Une fois les signaux transformés dans le domaine spectral, la densité spectrale croisée peut être calculée. La densité spectrale croisée S_{xy} est obtenue en faisant la moyenne dans le temps du module du produit du spectre du signal x et du conjugué complexe du spectre du signal y, comme montré dans l'équation 1.10. En pratique, cette estimation est généralement réalisée à l'aide de la méthode de Welch, qui consiste à appliquer une moyenne mobile aux signaux. Cette opération revient à utiliser un filtre à réponse impulsionnelle finie (*Finite Impulse Response*) (FIR) passe-bas. Une alternative à la méthode de Welch consiste à utiliser un filtre filtre à réponse impulsionnelle infinie (*Infinite Impulse Response*) (IIR) passe-bas (Same, Gandubert, Gleeton, Ivanov & Landry, 2020). À performance équivalente, les filtres IIR nécessitent beaucoup moins de calcul que l'approche de Welch, mais exigent davantage de précautions lors de leur conception, notamment pour assurer la stabilité du système. Ainsi, en pratique, la densité spectrale croisée S_{xy} peut être obtenue en appliquant un filtre passe-bas H dans le temps comme montré dans l'équation 1.11.

$$S_{xy}(f) = \lim_{t\to\infty} \frac{1}{t} |\mathbf{X}(f) \cdot \mathbf{Y}^*(f)|, \tag{1.10}$$

$$S_{xy}(f) = H(|\mathbf{X}(f) \cdot \mathbf{Y}^*(f)|) \tag{1.11}$$

La densité spectrale de puissance S_{xx} se calcule de la même manière que la densité spectrale croisée présentée dans l'équation 1.11, à la différence que le signal y est remplacé par le signal x. Enfin, la cohérence C_{xy} entre deux signaux est définie comme le rapport entre le carré du module de la densité spectrale croisée S_{xy} et le produit des densités spectrales de puissance S_{xx} et S_{yy} tel qu'illustré dans l'équation 1.12.

$$C_{xy}(f) = \frac{|S_{xy}(f)|^2}{S_{xx}(f) \cdot S_{yy}(f)} \tag{1.12}$$

Cette mesure fournit une indication du degré de corrélation fréquentielle entre les deux signaux, avec une valeur comprise entre 0 (absence de cohérence) et 1 (cohérence parfaite).

1.5.2 Détecteur d'activité vocale

La littérature portant sur les VAD est particulièrement abondante (Pertilä *et al.*, 2021). De manière générale, le processus de détection de la parole se compose de trois étapes : l'extraction de descripteurs, la prise de décision et, de façon optionnelle, une étape de lissage (Zhu, Zhang, Pei & Chen, 2023).

1.5.2.1 Extraction de descripteurs

L'extraction d'un ensemble de descripteurs permet de caractériser la voix en contraste avec le bruit. Idéalement, les descripteurs devraient prendre des valeurs distinctes selon qu'il y ait présence de voix ou seulement du bruit. Cependant, en pratique, la distribution des valeurs de la plupart des descripteurs présente un chevauchement entre les segments vocaux et non vocaux, en particulier en environnement bruité. L'utilisation conjointe de plusieurs descripteurs s'avère donc utile pour améliorer la fiabilité de la détection, en renforçant l'évidence d'activité vocale (Zhu *et al.*, 2023).

La cohérence binaurale, tout comme la cohérence entre un OEM et un IEM, peut être utilisée pour détecter la présence d'activité vocale. Le fonctionnement de cette approche est décrit dans

la sous-section 1.5.1. Comme mentionné précédemment, le calcul de la cohérence nécessite une transformation dans le domaine fréquentiel. De plus, la cohérence binaurale est sensible au bruit.

L'un des descripteurs les plus couramment utilisés est l'utilisation de coefficients cepstraux en fréquences de Mel (*Mel-Frequency Cepstral Coefficients*) (MFCCs) (Zhu *et al.*, 2023). Les MFCCs représentent la distribution de l'énergie spectrale d'un signal sonore dans le temps, sur une échelle de fréquences Mel. Ils sont obtenus à partir d'une transformation cosinus discrète appliquée à un spectre de fréquences non linéairement échelonné selon l'échelle Mel. Riches en information, les MFCCs sont largement utilisés dans des tâches telles que la reconnaissance vocale. Ils sont extraits à partir de la représentation fréquentielle du signal. Des implémentations standardisées pour leur extraction existent notamment dans MATLAB et Python. Toutefois, leur calcul peut être relativement coûteux en ressources. Ils sont souvent utilisé conjointement avec des algorithmes de prise de décision opaques comme le réseau de neurones profond (*Deep Neural Network*) (DNN) ou le HMM.

Similairement aux MFCCs, l'utilisation des coefficients prédictifs linéaires (*Linear Predictive Coefficients*) (LPCs) permet de recueillir une grande quantité d'informations sur un signal sonore (Zhu *et al.*, 2023). Les LPCs sont notamment utilisés pour la reconnaissance vocale. Contrairement aux MFCCs, ils sont extraits dans le domaine temporel. Les LPCs représentent un segment de signal à l'aide des coefficients d'un filtre à réponse finie, estimé à partir de l'autocorrélation du signal. Des algorithmes permettant leur extraction sont déjà disponibles dans MATLAB et Python. Les LPCs sont souvent utilisés conjointement avec des algorithmes de prise de décision opaques, tels qu'un DNN.

L'utilisation de centroïdes spectraux de sous-bandes normalisés (*Normalized Spectral Subband Centroids*) (NSSC) permet de déterminer l'équivalent du centre de gravité dans le domaine fréquentiel (Zhu *et al.*, 2023). Le NSSC correspond à la moyenne pondérée des fréquences d'un signal selon l'amplitude des différentes composantes fréquentielles. Ce descripteur est réputé pour sa résilience au bruit. Cependant, cette robustesse dépend de la plage fréquentielle du bruit : un bruit dont le spectre chevauche celui de la voix, comme un bruit d'usine, peut atténuer cette

propriété. Le NSSC est calculé à partir d'un signal dans le domaine fréquentiel. Des algorithmes pour extraire le NSSC sont déjà implémentés dans MATLAB et Python.

Le taux de passage par zéro (*Zero-Crossing Rate*) (ZCR) est calculé à partir du nombre de changements de signe dans une fenêtre temporelle (Zhu *et al.*, 2023). Des algorithmes permettant d'extraire le ZCR d'un signal sont déjà implémentés dans MATLAB et Python. Il s'agit d'un descripteur très simple à calculer. Cependant, le ZCR est particulièrement sensible à la présence de bruit. Ce descripteur est utilisé dans le domaine temporel.

Le SNR peut être estimé en s'appuyant sur le fait que les statistiques du bruit sont stationnaires sur une période plus longue que celles de la parole (Sohn, Kim & Sung, 1999). Ainsi, l'estimation du SNR est réalisée a priori. Cette estimation est ensuite comparée à un seuil prédéfini pour détecter la présence de bruit et d'activité vocale. L'estimation du bruit est mise à jour a posteriori. Ce descripteur est particulièrement efficace lorsque le bruit est statique.

Similairement au SNR, une décomposition en paquets d'ondelettes selon l'échelle de Mel peut être utilisée comme descripteur (Seris, Gargour & Laville, 2007). Le signal est converti à l'aide d'une décomposition en ondelettes, puis partitionné selon l'algorithme de Mallat. Les partitions normalisées sont regroupées et comparées entre elles à l'aide de la moyenne ou de la variance. La moyenne et/ou la variance peuvent être comparées à des seuils pour déterminer la présence de la voix. Ce descripteur est similaire au SNR, mais permet une meilleure précision dans les basses fréquences, où la voix est principalement contenue.

Le ratio entre différentes bandes de fréquences peut être utilisé comme descripteur pertinent (Lezzoum, Gagnon & Voix, 2014). Plus précisément, le rapport entre les énergies des bandes centrées entre 153 Hz et 1323 Hz est comparé à celui des bandes plus élevées (1323 Hz à 1944 Hz) et plus basses (15 Hz à 153 Hz). Ces ratios sont ensuite comparés à des seuils pour déterminer la présence de la voix. Ce descripteur a été développé pour offrir une bonne résilience au bruit.

Un comparatif des descripteurs listés, ainsi que de la conversion dans le domaine spectral, est présenté dans le tableau 1.3. À des fins de comparaison, le temps de calcul des différents descripteurs a été mesuré sur MATLAB, en réalisant 1000 itérations. Le temps médian est utilisé pour représenter la complexité des différents descripteurs. La cohérence entre deux signaux a été calculée en utilisant la méthode décrite dans la sous-section 1.5.1. Les fonctions fournies dans la librairie «Audio Toolbox» ont été utilisées pour calculer 16 MFCCs. Les fonctions fournies dans la librairie «Signal Modeling» ont été utilisées pour calculer 16 LPCs. Le NSSC a été calculé après conversion dans le domaine fréquentiel. Les fonctions fournies dans la librairie «Audio Toolbox» ont été utilisées pour calculer le ZCR. Le SNR a été calculé avec l'équation équation 1.7. Les ratios entre les différentes bandes de fréquences, notés $\frac{A_{f_{centrale}}}{A_{f_{haute}}}$ et $\frac{A_{f_{centrale}}}{A_{f_{basse}}}$, ont été implémentés et calculés à l'aide de MATLAB. La ligne incluant la FFT décrit la complexité de calcul liée à la transformation dans le domaine spectral grâce à MATLAB.

Tableau 1.3 Comparatif des descripteurs pour OVD

Descripteur	Complexité médiane(ms)	Résilience au bruit	Domaine
Cohérence OEM-OEM	0.29	Sensible	Spectral
Cohérence IEM-OEM	0.29	Résilient	Spectral
MFCC16	4.78	Résilient	Spectral
LPC16	1.04	Résilient	Temporel
NSSC	0.01	Variable	Spectral
ZCR	0.02	Sensible	Temporel
SNR	0.32	Variable	Interchangeable
$\frac{A_{f_{centrale}}}{A_{f_{haute}}}$, $\frac{A_{f_{centrale}}}{A_{f_{basse}}}$	0.21	Résilient	Interchangeable
FFT	0.17	-	Spectral

1.5.2.2 Prise de décision

Une fois les descripteurs extraits, une décision peut être prise afin de confirmer la présence de la voix. Différentes méthodes existent, nécessitant des durées d'entraînement variables et offrant des niveaux d'interprétabilité différents.

Une technique très interprétable pour la prise de décision est le schéma de prolongation. Une série de règles, parfois conditionnelles les unes aux autres, détermine la présence de la voix

(Lezzoum *et al.*, 2014; Seris *et al.*, 2007). Généralement, les descripteurs sont comparés à des seuils de détection, une valeur plus élevée indiquant la présence de la voix. De par leur nature, les schémas de prolongation sont très compréhensibles. Cependant, leurs performances sont généralement inférieures à celles des algorithmes plus opaques.

Le modèle de Markov caché (*Hidden Markov Model*) (HMM) peut estimer l'état d'un système dans une séquence à partir de descripteurs ainsi que de certaines probabilités apprises lors d'un entraînement, en utilisant l'algorithme de Viterbi (Pawełczyk *et al.*, 2011). Dans le cas d'un OVD, les états correspondent à la présence ou à l'absence de voix. Le modèle doit inclure la probabilité de l'état initial (quel état est actif au départ), les probabilités de transition (la probabilité de passer d'un état à un autre), ainsi que les probabilités d'émission (la probabilité d'observer un descripteur donné connaissant l'état du système). À partir des observations successives des descripteurs, l'algorithme calcule la séquence d'états la plus probable, ce qui permet de déterminer la présence ou l'absence de voix. Il existe déjà plusieurs implémentations dans différents langages de programmation, comme Matlab via la Statistics and Machine Learning Toolbox, C avec le Hidden Markov Model Toolkit, sur Python avec la librairie hmmlearn.

Les DNN offrent les meilleures performances en classification (Pertilä *et al.*, 2021). Un DNN prend une décision à partir d'un ensemble de couches de neurones. Un neurone est l'unité de base qui reçoit des informations, leur associe des poids, puis utilise une fonction d'activation pour transmettre un signal à un autre neurone ou à une sortie, comme illustré à la figure 1.2.

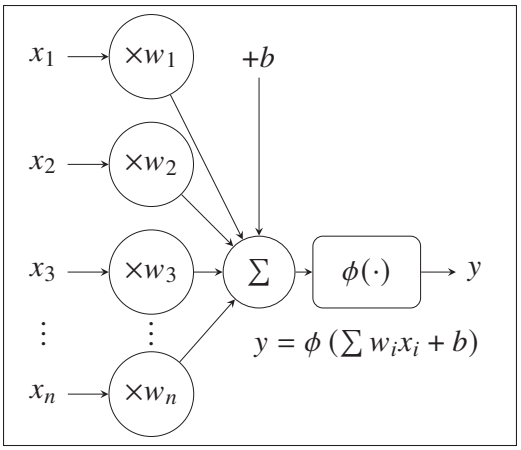

Figure 1.2 Représentation du fonctionnement d'une neurone

Les données brutes ou les descripteurs sont fournis à une couche de neurones d'entrée. Cette couche transmet ses signaux de sortie à une série de couches de neurones cachées. Enfin, une couche de sortie produit le résultat final de la classification, comme illustré dans la figure 1.3.

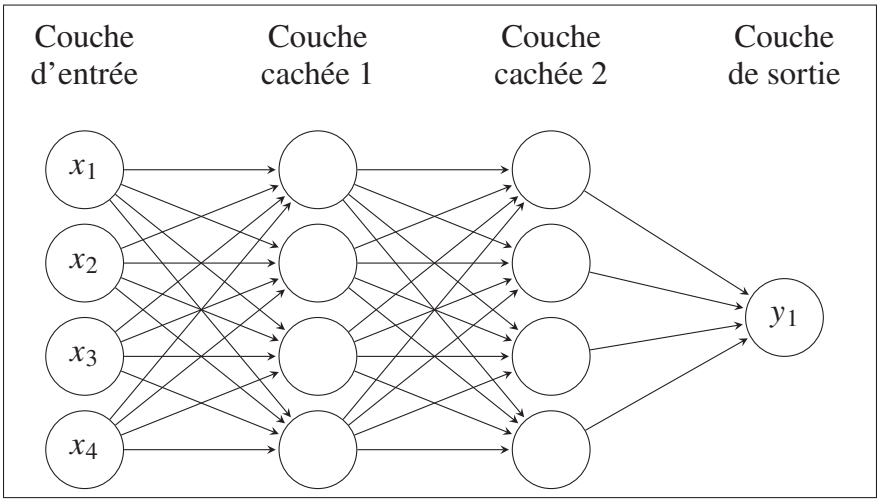

Figure 1.3 Représentation du fonctionnement d'un réseau de neurones

Si un seul neurone est utilisé avec une fonction d'activation sigmoïde, comme illustré dans l'équation 1.13, le modèle correspond à une régression logistique.

$$\phi(\cdot) = \frac{1}{1 + e^{-\sum w_i x_i + b}} \tag{1.13}$$

Les réseau de neurones convolutifs (*Convolutional Neural Network*) (CNN), une forme particulière de DNN, sont mieux adaptés au traitement en temps réel (Pertilä *et al.*, 2021). Contrairement aux DNN, les couches de neurones dans un CNN partagent les mêmes poids grâce à des filtres convolutifs. Cette architecture utilise des couches de plus en plus petites, permettant une extraction hiérarchique des caractéristiques du signal.

1.5.2.3 Techniques de lissage

Pour stabiliser la décision d'un VAD, des techniques de lissage peuvent être utilisées (Zhu *et al.*, 2023). Le but est de réduire les comportements erratiques de la détection.

La technique de lissage la plus simple consiste à filtrer les descripteurs ou la décision prise directement (Pang, 2017). Cette approche est facile à implémenter, mais elle peut introduire un délai dans la détection des transitions d'état, en fonction du temps de stabilisation du filtre.

Ensuite, les différents seuils utilisés peuvent avoir différentes valeurs pour l'activation et la désactivation (Sohn & Sung, 1998). Cette approche permet une transition rapide entre les états tout en évitant les fluctuations indésirables.

Similairement, les seuils d'activation peuvent être adaptés dynamiquement en fonction du niveau de bruit ambiant (Sohn *et al.*, 1999). Cette technique améliore les performances du VAD lorsque celui-ci repose sur des descripteurs sensibles au bruit, comme le SNR.

Une autre technique consiste à imposer un délai minimal entre l'activation et la désactivation du détecteur de voix (Lezzoum *et al.*, 2014). Cette approche s'intègre facilement dans un schéma de prolongation. Cependant, elle complique la détection précise des transitions entre deux locuteurs.

Lorsque la technique de prise de décision modélise explicitement les transitions d'état, comme c'est le cas avec un HMM, le lissage est intrinsèquement intégré au mécanisme de décision (Pertilä *et al.*, 2021).

1.5.3 Évaluation des performances

Les performances d'un OVD sont généralement évaluées à l'aide du ratio entre les bonnes et mauvaises détections (Lezzoum *et al.*, 2014; Pertilä *et al.*, 2021; Zhu *et al.*, 2023; Pawełczyk *et al.*, 2011). Ce ratio peut être exprimé à travers le taux de vrais positifs (TPR) et le taux de faux positifs (FPR). Ces deux indicateurs permettent de décrire la distribution des prédictions correctement identifiées et celles qui ne le sont pas. Dans le cas d'un OVD, un vrai positif correspond à un segment de voix correctement identifié. Un faux négatif correspond à un segment de voix classé à tort comme du bruit. Un faux positif correspond à un segment de bruit identifié à tort comme un segment de voix. Un vrai négatif correspond à un segment de bruit correctement identifié. Le TPR est défini comme le nombre de vrais positifs divisé par la somme des vrais positifs et des faux négatifs, comme illustré dans l'équation 1.14.

$$\text{TPR} = \frac{\text{Vrais positif}}{\text{Vrais positif} + \text{Faux négatif}} \tag{1.14}$$

Dans le cas d'un OVD, un taux de vrais positifs (TPR) élevé indique que peu de segments de voix sont manqués. Le taux de faux positifs (FPR) correspond au nombre de faux positifs divisé par la somme des faux positifs et des vrais négatifs, comme montré dans l'équation 1.15.

$$\text{FPR} = \frac{\text{Faux positif}}{\text{Faux positif} + \text{Vrais négatif}} \tag{1.15}$$

Dans le cas d'un OVD, un taux de faux positifs (FPR) élevé signifie qu'une grande quantité de bruit est incorrectement détectée comme de la voix.

1.5.4 Synthèse de la détection de la voix de l'utilisateur

Dans cette section, nous avons décrit comment les OVD détectent la présence de voix d'un utilisateur tout en excluant celles des autres. Cette distinction peut être réalisée à l'aide d'un accéléromètre ou par la cohérence entre deux microphones. Une fois qu'une activité est détectée, elle peut être confirmée comme étant de la voix grâce à un VAD. Un VAD s'appuie sur des descripteurs, certains plus adaptés au bruit que d'autres, pour alimenter un module de prise de décision. Les travaux portant sur des modules de prise de décision plus opaques dans leur fonctionnement rapportent de meilleures performances, mais sont plus difficiles à interpréter. Pour éviter les comportements erratiques, un OVD devrait intégrer une étape de lissage. Enfin, les performances d'un OVD sont souvent évaluées à l'aide du TPR et du FPR.

1.6 Modélisation de la voix en fonction de la distance

L'effort déployé par un locuteur pour produire sa voix est appelé effort vocal. Cet effort est influencé par de nombreux facteurs différents, tels que le niveau de bruit ambiant, la distance avec l'auditeur, la capacité du locuteur à s'entendre lui-même, le message qu'il souhaite transmettre, ses émotions, et bien d'autres.

Pour modéliser le phénomène d'effort vocal, Traunmüller & Eriksson (2000) ont enregistré 20 participants parlant à différentes distances dans un espace extérieur ouvert, face à un auditeur. Le groupe comprenait 6 hommes adultes, 6 femmes adultes, 4 garçons et 4 filles. Les distances variaient entre 0,3 m et 187,5 m. Le modèle proposé par Traunmüller & Eriksson (2000) montre clairement l'influence de la distance sur l'effort vocal. Ce modèle utilise un terme logarithmique et un terme exponentiel pour représenter l'effet de la distance d (en mètres), comme illustré dans l'équation 1.16 et l'équation 1.17. Le terme exponentiel corrige les limites du terme logarithmique, qui ne modélise pas correctement l'effort vocal à courte distance.

$$VEL \propto \log_2(d) \tag{1.16}$$

$$VEL \propto e^{1-n} | n(d) = \{(0.3, 1), (1.5, 2), (7.5, 3), (37.5, 4), (187.5, 5)\} \tag{1.17}$$

La vitesse du vent a également été prise en compte dans le modèle, même si l'expérience initiale ne prévoyait pas d'intégrer cet élément. Il est probable que le vent ait généré du bruit significatif durant les enregistrements. Le modèle repose ensuite sur différents facteurs, tels que l'âge, le genre, ainsi que des constantes spécifiques à chaque participant. Il établit une relation entre ces facteurs et plusieurs paramètres vocaux : le SPL de la voix, la fréquence fondamentale, les formants, ainsi que les pauses entre les mots prononcés. Ce modèle ne prend pas en compte directement le niveau de bruit ambiant, mais il couvre une large plage de distances entre locuteur et auditeur.

Pelegrín-García, Smits, Brunskog & Jeong (2011) ont modélisé l'effort vocal de 13 hommes dans différentes pièces : salle anéchoïque, corridor, salle réverbérante et salle de classe. Les facteurs influençant l'effort vocal étaient l'acoustique de la pièce, la distance entre le locuteur et l'auditeur, ainsi que le niveau de bruit ambiant. Le SPL de la voix et la fréquence fondamentale ont été prédits à l'aide d'un modèle intégrant ces différents facteurs. Ce modèle a été développé dans des environnements où le bruit ambiant était relativement faible, inférieur à 45 dBA.

Bouserhal, Bockstael, MacDonald, Falk & Voix (2017a) se sont intéressés à l'effort vocal lors du port de HPD en fonction de la distance et de la quantité de bruit, avec 12 participants (2 femmes et 10 hommes). Le modèle proposé permet de prédire le SPL de la voix. Il repose sur une relation logarithmique avec la distance, une relation linéaire avec la quantité de bruit, le SPL de la voix du locuteur à 10 m d'un auditeur, ainsi que sur l'interaction entre ces différents termes. Ce modèle intègre le port de HPD, la présence de bruit industriel et différentes distances, ce qui le rend plus adapté que les autres pour modéliser l'effort vocal dans un environnement industriel.

1.7 Spatialisation du son

La spatialisation du son peut être réalisée à l'aide d'une fonction de transfert liée à la tête (*Head-Related Transfer Function*) (HRTF) (Niermann, 2015). La HRTF inclut le comportement de l'IID et de l'ITD dans un filtre. Une HRTF est un ensemble de paires de filtres pour l'oreille

gauche et l'oreille droite, classées selon la provenance du son en fonction de l'azimut et de l'élévation. La figure 1.4 illustre un son se dirigeant vers une personne située à l'origine, cette personne regardant dans la direction de l'axe des abscisses. Le son est représenté par un vecteur. L'azimut correspond à l'angle entre l'axe des abscisses et la projection du vecteur sur le plan XY (représentée en pointillés). L'élévation est l'angle entre la projection sur le plan XY et le vecteur.

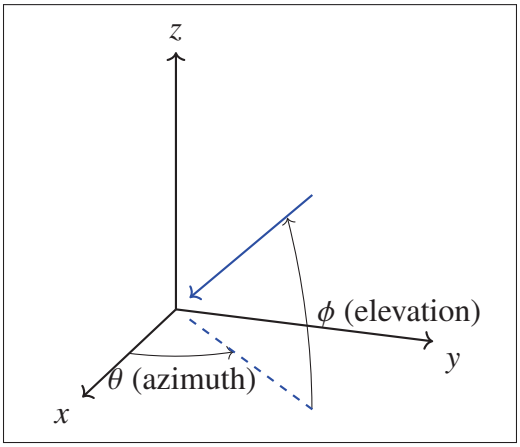

Figure 1.4 Représentation de l'azimuth et de l'élévation d'un son

En appliquant une paire de filtres sur un signal transmis électroniquement, il est possible de créer l'illusion d'une direction spécifique. Les HRTFs sont relativement uniques à chaque individu, car elles dépendent de la géométrie corporelle (Pelzer *et al.*, 2020). Les HRTFs possèdent un équivalent pour la production de la voix. Cet équivalent est modélisé par une fonction de radiation de la parole, qui décrit le champ de SPL autour du corps du locuteur en relation avec la bouche (Bellows & Leishman, 2022; Dunn & Farnsworth, 1939). Ce champ comprend l'azimut et l'élévation, comme pour une HRTF, mais aussi un rayon mesuré à partir de la bouche.

Ainsi, pour une spatialisation complète de la voix d'un locuteur vers un auditeur, le profil fréquentiel et l'intensité de la voix doivent être modifiés par une fonction de radiation de la parole, adaptés en fonction de la distance — comme mentionné dans la section 1.1 — puis filtrés par une HRTF.

1.8 Synthèse de la revue de littérature

Cette revue de littérature a permis de présenter les fondements nécessaires à la modélisation du son, de la voix et de l'écoute, comme détaillés dans la section 1.1. La section 1.2 introduit la protection passive à l'aide des Hearing Protection Device (HPD), en abordant les différents types existants. Il y est montré que le bruit nuit à l'intelligibilité, ce qui souligne la nécessité de solutions efficaces, tant pour les personnes ayant une audition normale que pour celles atteintes de pertes auditives. La section 1.3 souligne l'intérêt de diverses industries pour la spatialisation du son et l'intelligibilité vocale, deux facteurs ayant un impact direct sur la charge de travail et la conscience situationnelle. Dans plusieurs cas, le port de bouchons dotés d'ANC apparaît comme une solution prometteuse. La section 1.4 présente plusieurs exemples d'implémentations d'AHPDs intégrant un système de communication fondé sur l'usage d'IEMs. Les algorithmes mis en œuvre dans ces dispositifs incluent notamment des filtres antibruit, la détection d'activité vocale et des systèmes d'ANC. La section 1.5 passe en revue les différents algorithmes liés aux OVD, en mettant en lumière leur complexité de calcul, leur opacité, et le matériel nécessaire. La section 1.6 se concentre sur la modélisation de l'effort vocal, influencé par le bruit ambiant, la distance, le port de HPDs et des facteurs individuels. Enfin, la section 1.7 explique comment spatialiser le son à l'aide d'une HRTF et de fonctions de radiation de la parole. Dans l'ensemble, cette revue de littérature permet de mieux comprendre l'intérêt de RAVE et les principes qui sous-tendent la conception de son banc d'essai.

CHAPTER 2

ENABLING PERSONALIZED COMMUNICATION WITH ADVANCED HEARING PROTECTION DEVICES: INTEGRATING AND EVALUATING CONCEPTS OF A RADIO ACOUSTICAL VIRTUAL ENVIRONMENT

Gabriel Ouaknine-Beaulieu[1,3] , Xinyi Zhang[1, 3] , Charles Pillet[1] , Jérémie Voix[2] , Rachel Bouserhal[1, 3] , Pascal Giard[1]

[1] Département de Génie Électrique, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[2] Département de Génie Mécanique, École de Technologie Supérieure,
1100 Notre-Dame Ouest, Montréal, Québec, Canada H3C 1K3
[3] Centre interdisciplinaire de recherche en musique, médias et technologie,
527 Sherbrooke Ouest, Montréal, Québec, Canada H3A 1E3

Article soumis à la revue « Journal of the Acoustical Society of America » en juillet 2025.

2.1 Abstract

To facilitate communication between workers in noisy environments, active systems featuring interpersonal radio communication may be used despite the fact that these systems do not enable users to dynamically address specific individuals or perceive the directionality of speech. A promising approach to these challenges is an emerging technology called Radio Acoustical Virtual Environment (RAVE). RAVE aims to improve communication in two ways: (1) allowing users to dynamically address specific individuals based on vocal effort, and (2) transforming speech signals to convey spatial directionality of its origin. To explore its potential, a mockup version of RAVE incorporating key algorithms was developed and evaluated with 18 participants. The evaluation demonstrated RAVE's potential to improve communication. Additionally, the results offer valuable insights into user interaction and inform future design decisions for Active Hearing Protection Device (AHPD) with integrated communication systems.

2.2 Introduction

To prevent the harmful effects of noisy environments, hearing protection is needed. While passive Hearing Protection Device (HPD) effectively reduce the level of harmful noise, they also degrade

speech perception, making communication more difficult (Casali & Horylev, 1987). For those with hearing loss, ambient noise further diminishes intelligibility, hindering clear workplace communication—a critical factor for safety and productivity (Giguère *et al.*, 2010). Over the past three decades, Active Hearing Protection Device (AHPD) such as Tactical Communications and Protective Systems (TCAPS) have evolved to improve communication and situational awareness in various environments, such as the military (Casali, 2010; Abel *et al.*, 2007; Brammer *et al.*, 2014a) and aviation (Niermann, 2015; Ribera *et al.*, 2004; Casto & Casali, 2013). In contrast, industrial environments have seen limited advancements in AHPD specifically designed to enhance communication between workers (Pawełczyk *et al.*, 2011). Existing research primarily focuses on Active Noise Control (ANC) and alarm signal detection (Laroche *et al.*, 2022; Paliwal, Lyons & Wójcicki, 2010). Although some research addresses situational awareness and cognitive workload reduction (Casto & Casali, 2013; Abel *et al.*, 2007; Niermann, 2015; Lee & Casali, 2017), there are no existing practical solutions for industrial environment. Additionally, conventional AHPD lack the ability to identify the intended listener, often relaying irrelevant speech and causing miscommunication. One promising approach for communication in noisy industrial environments is a Radio Acoustical Virtual Environment (RAVE) (Bou Serhal *et al.*, 2013). RAVE is a recently proposed technological concept for improving communication in industrial environments. Unlike conventional communication systems, which either broadcast speech to all users or require the selection of specific channels, RAVE aims to dynamically target speech transmission to only intended listeners, the specific person or group that is aimed by the talker, instead of everyone on the same radio channel. The intended listeners are all users within a communication radius. The communication radius is determined by an estimation of the speaker's vocal effort and the surrounding noise level. Hence, RAVE has the potential to allow communication to be more context-sensitive and intuitive, enhancing users' experience on both the emitting and receiving sides. Additionally, RAVE wants to eliminate the need for traditional Push-To-Talk (PTT) systems, reducing cognitive load, potentially leading to more productivity. Additionally, like military and aviation settings, RAVE attempts to improve situational awareness. It mimics speech localization at the listener's end (reception) through spatialized audio filtering, enabling better directional perception. The objective is to create a sense of directionality, which

allows users to localate talkers and engage in “targeted communication”. RAVE is also designed to improve the speech-to-noise ratio, addressing common limitations of traditional AHPD; conventional systems, such as those using boom microphones, require precise alignment with the user’s mouth to effectively reject noise and are incompatible with other personal protective equipment like respirators. Using an advanced earpiece featuring an In-Ear Microphone (IEM) and Outer-Ear Microphone (OEM), as well as a miniature loudspeaker, RAVE aims to overcome these challenges by capturing speech directly from within the occluded ear canal, enhancing the speech-to-noise ratio. RAVE is designed to use both the IEM and OEM, which allows to filter and refine the audio signal, providing a clearer communication experience (Bouserhal, Falk & Voix, 2017b). In this paper, a mockup implementation of RAVE is presented, addressing gaps in previous research by integrating and linking the functional components of this emerging technology that have been so far explored in isolation. The mockup implementation of RAVE provides valuable insights into the interplay between its various technological components. Moreover, this paper represents a pioneering contribution by presenting the first ever participant-based evaluation of a RAVE mockup in dynamic scenarios. Participants tested the mockup implementation of RAVE and of PTT, providing feedback through surveys evaluating ease-of-use, communication efficiency, and the overall realism of the mockup. The results are used to highlight both the advantages and limitations of the emerging technology, offering an initial conceptualization of how RAVE could function in the real world. Ultimately, the goal is to inform future design decisions, hardware technological choices, and subsequent iterations of RAVE, as well as other AHPD communications systems.

An overview of the different technologies needed for RAVE is presented in section 2.3. Then, section 2.4 describes the methodology used to test the mockup implementation of RAVE with participants. The implementation itself is described in subsection 2.4.4. Participants’ feedback are presented in section 2.5, while the performance of RAVE’s algorithms is discussed in section 2.6. The discussion and conclusions are presented in section 2.7 and section 2.8 respectively.

2.3 Background

AHPD equipped with personal radio communication systems rely on various algorithms to function effectively. Only a few modern communication systems specifically designed for industrial noise environment tackle IEM-related challenges as discussed in subsection 2.3.1 (Pawełczyk *et al.*, 2011; Westerlund *et al.*, 2005). Similar to those other systems, RAVE relies on IEMs and algorithms for Own Voice Detection (OVD), Intended Listener Detection (ILD), and speech enhancement. The existing solutions to these challenges are reviewed in subsection 2.3.2, subsection 2.3.3, and subsection 2.3.4, respectively.

2.3.1 Communication Systems Based on In-Ear Microphone

Two prior research works describing the challenges associated with communication systems using IEMs in industrial environments are described in this section. The use of IEMs induces specific challenges such as detecting the voice of the wearer, reducing noise picked-up by the IEM, improving the bandwidth of the speech picked-up, and performing ANC the active cancellation of the sound pressure at the ear.

The first work describes a communication system combining earmuffs and IEMs that offers ANC and communication function (Westerlund *et al.*, 2005). ANC is achieved by an analog feedback system based on a reference microphone and a loudspeaker in the cups of the earmuffs. The Voice Activity Detector (VAD) is based on a Hidden Markov Model (HMM) with a 76.7 % True Positive Rate (TPR). Even with ANC, this setup requires extra noise reduction, which is achieved with spectral subtraction of the captured speech signal. Westerlund *et al.* (2005) describe IEM speech as a low-pass filtered version of speech from the mouth. The authors improve speech intelligibility and quality with a simple high-pass filter.

The second work describes a communication system for grouped users with earplugs equipped with IEMs that offers ANC and speech enhancement (Pawełczyk *et al.*, 2011). This work focuses more on the real-time aspect, which requires minimizing computational complexity. It is briefly described that users in the same zone are given a radio communication channel in a “hot-spot”

style configuration. Up to three simultaneous talkers could talk with only the users located in the same zone. ANC is achieved with a feed-forward control structure. Noise reduction is achieved with adaptive Least Mean Square (LMS) filtering, where the reference comes from the OEM recording only the noise. The implementation is said to require a VAD, but the algorithm itself is not described by the authors.

2.3.2 Own Voice Detection

For RAVE to function properly, an OVD is required that can detect when the user is speaking, without being activated by background noise or speech from others. Modern OVD algorithms distinguish users' voice activity using what this paper refers to as a Wearer Induced Disturbances Detector (WIDsD). Modern OVD algorithms typically detect speech using a VAD.

Discriminating between the wearer's voice and other close-by voices can be achieved through accelerometers or by comparing the right and left channels of the OEMs (Pertilä *et al.*, 2021). Since the binaural OEM signal approach was initially developed for hearing aids, its performance can be compromised in environments with elevated noise levels. In in-ear dosimetry applications, Wearer Induced Disturbances (WIDs), such as coughing, microphonics, and speech, have been detected using the coherence between the IEM and OEM signals (Bonnet *et al.*, 2019). Using the coherence method is advantageous because it requires less hardware.

VAD extracts features from speech, makes decisions based on those features, and may apply smoothing techniques to minimize classification error rates (Zhu *et al.*, 2023). Traditional features for VAD include Speech-to-Noise Ratio (SNR) estimation (Sohn *et al.*, 1999), Mel-Frequency Cepstral Coefficients (MFCC), Linear Predictive Coding (LPC), normalized spectral subband centroids, and zero-crossing rate (Zhu *et al.*, 2023). Additionally, the coherence between two microphones can aid in detecting voice activity even in the presence of coherent interference (Kim & Cho, 2006). The effectiveness of these features depends on the noise environment, and more complex features such as MFCC and LPC can take more time to compute. Some features are specially designed for industrial settings, like Mel-scale wavelet packets (Seris *et al.*,

2007). A simpler alternative feature for industrial settings uses the ratio between frequency bands (Lezzoum *et al.*, 2014).

Once features are extracted, various decision-making approaches exist for OVD. Traditional methods include hangover schemes, as well as HMM or Switching Kalman Filters (Lezzoum *et al.*, 2014; Westerlund *et al.*, 2005; Fujimoto & Ishizuka, 2008) approaches. More recent advancements leverage Deep Neural Network (DNN), which offer superior performance (Pertilä *et al.*, 2021). In particular, compact neural networks are better suited for real-time applications (Sehgal & Kehtarnavaz, 2018). DNN achieve the highest performance among current methods; however, they are computationally intensive, require extensive training on annotated datasets, and their decision-making processes are often difficult to interpret. Similarly, HMM, and Switching Kalman Filters also necessitate training and exhibit limited interpretability, though they are better suited for real-time applications. By contrast, hangover schemes offer greater interpretability and do not require prior training, but they generally underperform compared to data-driven approaches.

Decision smoothing techniques are often integrated within algorithms, as seen in HMM and hangover schemes, but can also be implemented separately using hysteresis thresholding, noise-adaptive thresholding (Sohn & Sung, 1998; Sohn *et al.*, 1999), or feature averaging for enhanced robustness (Pang, 2017).

2.3.3 Intended Listener Detection

The concept of RAVE is based on the talker's vocal effort, which has been modeled in various ways. Vocal effort can only be inferred from the distance between the talker and the listener, and from variations in voice parameters (Fux, 2012). The method proposed by Traunmüller & Eriksson (2000) estimates vocal effort using voice Sound Pressure Level (SPL) and frequency profile, taking into account factors such as distance, gender, age, and wind speed. The method proposed by Pelegrín-García *et al.* (2011) incorporates voice SPL, frequency profile, and phonation time ratio, considering distance, room acoustics, and ambient noise SPL. Unlike these models, which

assume non-occluded ears, the method proposed by Bouserhal *et al.* (2017a) estimates vocal effort under occluded conditions using voice SPL, distance, ambient noise SPL, and voice loudness. This last approach is particularly relevant for RAVE, as it accounts for ear canal occlusion, making it more suitable for realistic vocal effort estimation in industrial settings, where hearing protectors, especially earplugs, may already be worn.

2.3.4 Speech Enhancement for In-Ear Microphones

Once the talker and the intended listeners have been identified, the talker's speech needs to be enhanced before being played in the intended listeners' ears. Speech enhancement in RAVE consists of three key steps: 1) residual noise reduction to remove ambient noise passing through the HPD and WIDs, 2) Bandwidth Extension (BWE) to compensate for frequency losses due to bone and tissue conduction, and 3) spatialization using Head-Related Transfer Function (HRTF) and the talker's voice directivity pattern.

Despite the passive attenuation of the HPD, in high noise levels, the residual noise may still degrade speech quality (Bouserhal *et al.*, 2017b). To improve speech quality, noise reduction techniques can be applied to the captured IEM signal. Spectral subtraction offers a viable noise-reduction method, particularly with newer optimizations that mitigate musical noise and reduce dependency on VAD (Westerlund *et al.*, 2005). Although it operates in the spectral domain—making it computationally demanding—spectral subtraction can be highly effective due to its potential for a negligible transition band, thereby limiting signal distortion. Multi-band adaptive gain control presents a simpler option, eliminating the need for VAD (Lezzoum, Gagnon & Voix, 2016a). However, it is less selective, and despite the time-frequency approach used, it struggles to suppress strong noise components that overlap with the speech frequency range, thus limiting its overall effectiveness. Normalized LMS adaptive filtering, in combination with IEM and OEM, can effectively suppress loud noise (Bouserhal *et al.*, 2017b). While this approach operates in the time domain and can provide complete noise removal, it requires an OEM and further relies on the use of a VAD to assess when to update the coefficients of the adaptive filter.

Due to the limited bandwidth of IEM speech, 0-2 kHz, causing the captured speech to sound “boomy" and reduce perceived speech quality, BWE techniques should be applied (Bouserhal *et al.*, 2017b). Basic high-pass filtering introduces no delay (Westerlund *et al.*, 2005), while more advanced methods–such as up sampling combined with filtering and signal processing (Bouserhal *et al.*, 2017b) or neural networks (Park, Shin & Shin, 2019)–can be used for bandwidth reconstruction. While high-pass filtering is efficient, it has not been shown to improve perceived speech quality to the same extent as more sophisticated methods. In addition to requiring training, neural networks did not report performance as good as up sampling combined with filtering and signal processing.

Finally, to enhance spatial awareness, HRTF and voice directivity patterns can be applied on speech. HRTF model how sound is altered by the listener’s morphology based on sound incidence (level, frequency content, and directivity). Utilizing an HRTF can create a sense of spatial direction in sound. Some research even explored methods to match HRTF to individual anatomical characteristics (Pelzer *et al.*, 2020). Different HRTF databases already exist, such as the HUTUBS database (Brinkmann *et al.*, 2019) or the RIEC database (Majdak *et al.*, 2013). Conversely, the voice directivity pattern models how sound propagates from the mouth to surrounding directions (Bellows & Leishman, 2022; Dunn & Farnsworth, 1939).

2.3.5 Radio Acoustical Virtual Environment Signal Processing

The complete Digital Signal Processing (DSP) audio path needed to implement RAVE is illustrated in Figure 2.1. First, speech signal is recorded from the talker’s earpiece using both the IEMs and OEMs. Depending on the chosen DSP technique, the signal may subsequently be transformed into the spectral domain. Then, the OVD and ILD algorithms identify the talkers and intended listeners, respectively. Next, the processed speech signal is transmitted from the identified talkers to the appropriate listeners. Finally, the signal is enhanced, converted back to the time domain, and played through the listener’s earpiece. Reversing the order of signal enhancement and OVD—applying enhancement first—would increase average processing requirements but reduce the need for a highly noise-robust OVD.

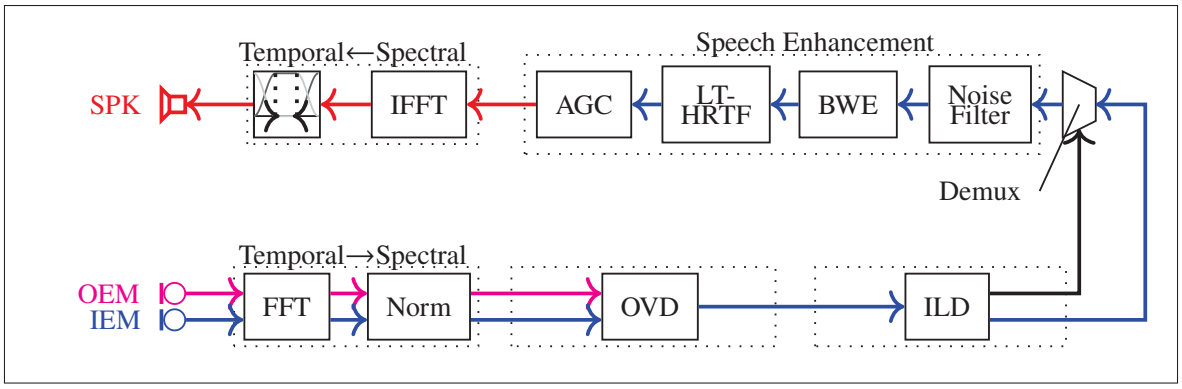

Figure 2.1 Implementation of RAVE with audio from the talker's IEM and OEM to the intended listener's loudspeaker (SPK)

The *blue* line depicts the signal picked up by IEM, the *magenta* line depicts the signal picked up by OEM, the *black* line represents the distance information and the *red* line the downlink audio signal. FFT and IFFT denote the Fast Fourier Transform and its inverse operation. Norm represents signal normalization. The OVD, ILD and BWE boxes represent the eponym algorithms for own voice detection, intended listener detection and speech bandwidth extension. The LT-HRTF represent the combination of HRTF and voice directivity patterns for sound spatialization. The AGC refers to Automatic Gain Control. The overlapping windows are represented by the graph box

2.4 Methods

2.4.1 Participants

Participants were invited to take part in the experiment through an email incitation approved by the Comité d'éthique pour la recherche, the internal review board of the École de technologie supérieure (H20231110). Prior to the data collection, their hearing ability was evaluated with a pure tone audiometric screening test, following the Hughson Westlake threshold tracking procedure, and an otoscopic inspection was conducted to verify that participants were eligible for the study. Each participant had a hearing threshold lower than 25 dBHL at 500 Hz, 1 kHz, 2 kHz, 3 kHz, 4 kHz, 6 kHz, and 8 kHz. All participants confirmed that they were at least 18 years of age, fluent in English, and did not self-identify as having hypersensitivity in the ear canal. Additionally, they reported no current ear infections or inflammation, no history of ear surgery, no diagnosed voice disorders, and no difficulty in locating the direction of sound. All

participants gave their consent to participate in the experiments and could withdraw their consent at any time.

2.4.2 Experimental Protocol Procedure

Upon arrival, participants were given a brief orientation outlining the data collection procedure and an explanation of each mockup operation. Each participant was then equipped with a hard hat and a belt containing wireless transmitters and a receiver. Motion-capture reflectors were affixed to the top of the hard hats to track each participant's localization in the room. Clean earplugs were distributed, and the organizer ensured they were properly inserted. A broadband white noise was then played through external loudspeakers, and earplug attenuation was calculated using the IEM and the OEM signals and the F-MIRE method (Voix & Laville, 2009). Once a proper fit was confirmed, each participant was directed to their assigned desk for data collection. A construction set, a manual stating the 24 instructions to build the set, and a list of all block identification numbers were on each desk. The construction sets did not vary in difficulty, as they were all designed for 8-year-olds. All construction sets had some missing blocks that were placed on a different participant's desk to provoke communication between participants during the data collection.

The data collection was divided into scenarios, each lasting 10 minutes. An overview of the experimental setup is depicted in Figure 2.2. During each scenario, participants—represented as green circles—faced outward from the center. Participants were required to communicate with one another because certain blocks were missing from their own sets and were present in other participants' sets. If a participant was unable to follow the instructions due to a missing block, they were instructed to ask the others, only one other participant at a time while facing them directly, if they had it. If the contacted participant possessed the requested block, they were expected to deliver it to the requesting participant. At the midpoint of each scenario, participants were instructed to switch positions with their desks, as illustrated in Figure 2.2 and Figure 2.3. After 10 minutes, the scenario concluded, and participants completed a survey questioning. The survey's questions focused on the ease of use of the communication system, the efficiency

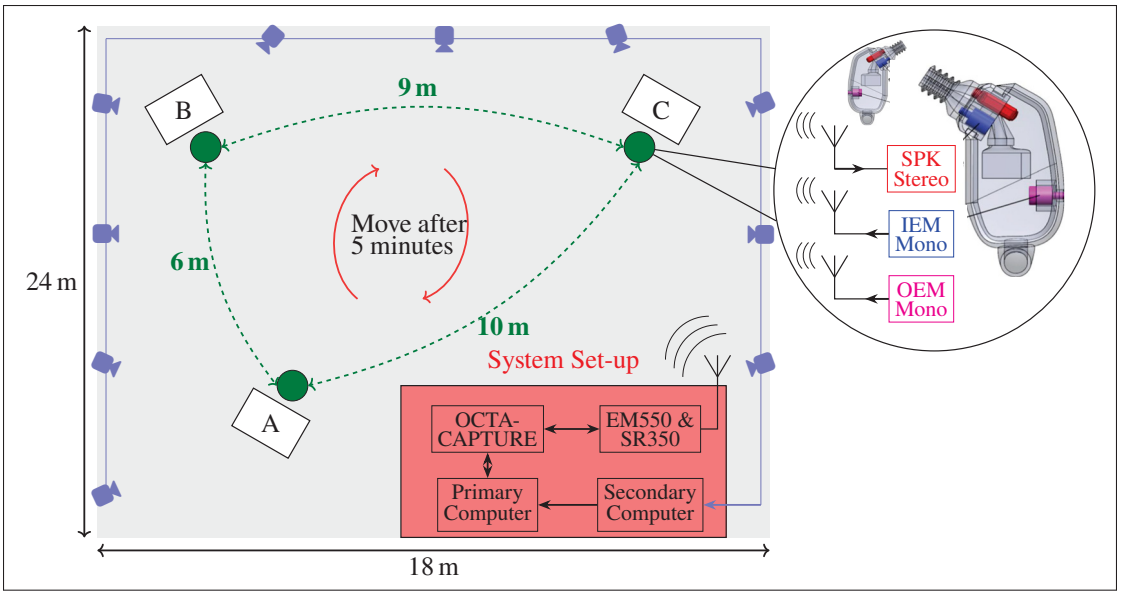

Figure 2.2 Experimental setup using PTT and RAVE mockups

The grey box represents the MultiMedia Room (MMR) along with its dimensions. Green circles indicate participant positions. The red box shows the table containing the audio interface, receivers, transmitters, and computers. The zoomed-in inset illustrates the earpieces, receivers, and transmitters worn by the participants

of communication, and the realism of the mockup. Participants were also invited to provide comments either written or verbally.

A total of 4 scenarios have been tested: 2 using a RAVE mockup and 2 using a PTT system mockup. Moreover, both PTT and RAVE mockups were tested with faint white and broadband factory noises. The experiment always started with white noise; however, half of the groups started with the PTT mockup, while the other half started with the RAVE mockup. The type of noise changes after each scenario. The industrial noise consisted of a 75 dBA factory noise (Varga & J.M.Steeneken, 1993) with the spectral envelope depicted in Figure 2.4. Once all scenarios were completed, participants removed their radio belts and hard hats. Each participant was thanked with a compensation of $20 or a pair of musician earplugs (ETY 20 plugs, Etymotic).

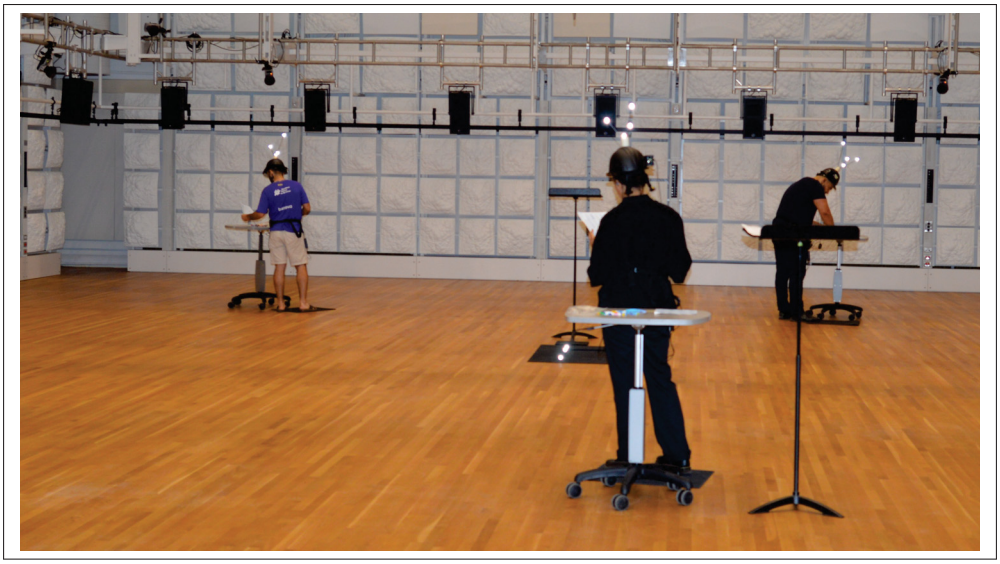

Figure 2.3 Experimental setup with participants

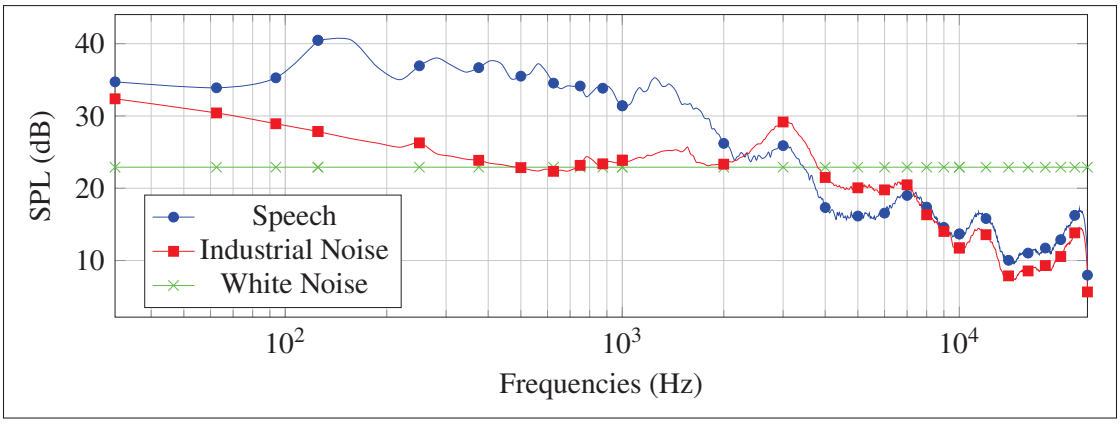

Figure 2.4 Spectral envelopes

Spectral envelope of the industrial noise with 74.12 dBA SPL (red curve), white noise with an overall level equivalent to the 74.12 dBA SPL of the industrial noise (green curve), and IEM speech from a participant (blue curve) with 84.70 dBA SPL. The *x*-axis represents the frequencies on a logarithmic scale

2.4.3 Materials

To conduct the data collection, an extensive set of materials is required, with some depicted in Figure 2.2 and in Figure 2.5. A main computer ran the DSP, exchanged audio signals with an

Figure 2.5 Experimental material: two computers, three stereo transmitters, and six mono receivers

audio interface (Roland, OCTA-CAPTURE) at 24-bit precision and 48 kHz sampling rate. The audio interface was connected to three stereo transmitters (Sennheiser, SR350) and six mono receivers (Sennheiser, EM550), ensuring synchronized audio delivery. Each participant was equipped with a stereo receiver (Sennheiser, EK300) attached to a belt, delivering separate audio signals to each ear. Two mono transmitters (Sennheiser, SK500) were also attached to each belt–one dedicated to the OEM and the other to the IEM. In PTT scenarios, participants activated the system using a WiFi-enabled button built with a micro-controller (Espressif, ESP32). The state of the PTT button was transmitted via a WiFi UDP connection with a refresh rate of one second to the main computer. The participants wore hard hats fitted with 30 mm motion-capture reflectors to enable precise tracking of head movements. The secondary computer processed video feeds from motion-capture cameras (Qualisys, Oqus 310 and 400), calculated the position and orientation of participants using Qualisys Track Manager (Qualisys, Göteborg, Sweden), and sent this data to the main computer at 46 Hz via a wired TCP connection. The motion-capture

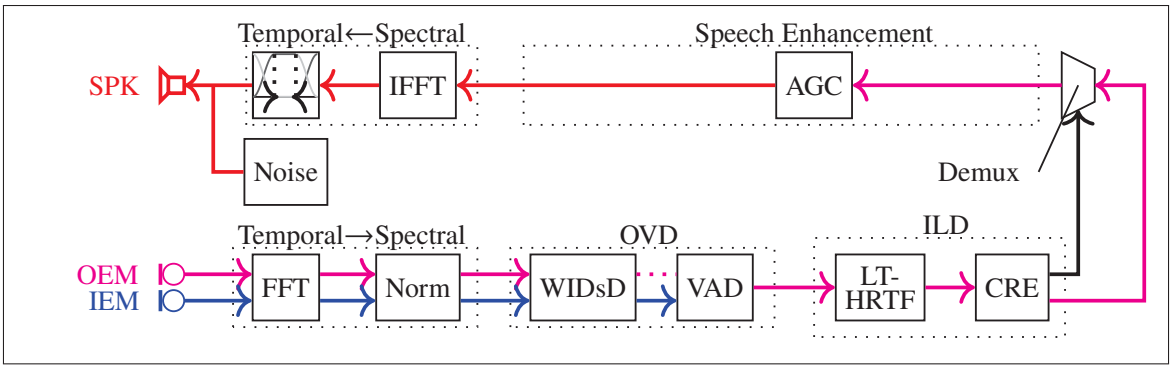

Figure 2.6 Mockup implementation of RAVE with audio from the talker's IEM and OEM to the intended listener's loudspeaker (SPK)

The *blue* line depicts the signal picked up by IEM, the *magenta* line depicts the signal picked up by OEM, the *black* line represents the distance information and the *red* line the downlink audio signal. FFT and IFFT denote the Fast Fourier Transform and its inverse function. Norm represents audio signal normalization. The WIDsD and VAD boxes represent the eponym algorithms to detect wearer's induced disturbances and voice activity detection. The LT-HRTF represents the combination of HRTF and voice directivity patterns for sound spatialization. The CRE box represents the decision for intended listener based on vocal effort. The AGC refers to Automatic Gain Control. The overlapping windows are represented by the graph box

cameras were installed all around the MMR room and their 9 m range determined the usable space for the experiment.

2.4.4 Implementation

Both mockups use MATLAB R2022b (MathWorks, Natick, Massachusetts, USA) to carry out the DSP, with a focus on reducing complexity and minimizing algorithmic errors for real-time constraints, while integrating important functionalities such as OVD (subsubsection 2.4.4.3), ILD (subsubsection 2.4.4.5), and sound spatialization (subsubsection 2.4.4.4). The speech processing pipeline captures audio, converts it to the spectral domain, performs OVD and ILD, adjusts the gain appropriately for the listener, and then transforms the signal back to the time domain. Finally, it adds the simulated residual ambient noise (either white or factory) and plays the signal through the earpiece loudspeaker of the intended listener. Figure 2.6 illustrates this pipeline.

2.4.4.1 Mockup Simplifications and Constraints

The mockups are simplified to respect the available computational power for real-time applications. The first simplification is that they operate in half-duplex mode, allowing only one talker to speak to the others at a time. Full-duplex communication systems are prone to audio feedback and require specialized control mechanisms, which exceed the available computational power. Similarly, instead of implementing ANC to limit the residual ambient noise picked up by the IEM, the experiment is conducted in a quiet room and a synthetic ambient noise is directly added to the audio played in each participant's ear. This allows precise control over noise levels while minimizing complexity. Because the residual ambient noise is introduced directly into the ear, speech enhancement is minimized by using the clean speech picked up by the OEMs. Therefore, the listener receives a wideband clean signal, eliminating the need for IEM speech enhancement algorithms such as noise reduction and BWE, which can be computationally exhaustive. OVD is still performed using the IEMs, while the other algorithms rely on OEMs.

Since the three participants are relatively close to each other, speech is spatialized (as detailed in subsubsection 2.4.4.4 before estimating the communication radius to amplify the distinction between possible intended listeners.

Finally, the RAVE mockup determines (as detailed in subsection 2.4.3 participants' head positions using motion capture. A commercialized implementation would instead rely on a combination of an inertial measurement unit and a ranging algorithm.

2.4.4.2 Recording Audio

To process audio in real-time within MATLAB, digital signal processing is performed on a short window in a continuous while loop. To minimize latency, the loops are designed to use the smallest window size that still allows enough time to process the whole DSP within the frame's duration. Window too short would not allow enough time to process the whole DSP and result in a critical fail error. Each loop begins by recording the audio from both the OEM and the IEM for the last $T_{\text{window}} = 21.3\,\text{ms}$. This audio is then concatenated with the previous 10.6 ms of data

to form a more suitable window length. When reconstructing speech from a spectral signal, the ideal window duration typically falls between 15 ms and 35 ms (Paliwal *et al.*, 2010) hence, the window is within that range. The audio is then transformed into the frequency domain using the Fast Fourier transform (FFT) algorithm, and the overall level is expressed as SPL in dB.

2.4.4.3 Own Voice Detection

For participants not receiving speech, the frequency-domain signals from their corresponding OEM and IEM are analyzed using the OVD algorithm. The OVD consists of two modules: a WIDsD and a VAD module. The WIDsD operates by calculating the coherence function between the OEM and IEM signals as (Bonnet *et al.*, 2019):

$$\gamma^2(f) = \frac{|S_{OI}(f)|^2}{S_{OO}(f)S_{II}(f)}, \tag{2.1}$$

where S_{OI} is the cross-spectral density between the OEM and IEM, S_{II} is the power spectral density of the IEM, and S_{OO} is the power spectral density of the OEM. Traditional coherence libraries are computationally intensive and are thus not suitable for real-time systems. As a consequence, an optimized version of the Welch algorithm has been implemented using Infinite Impulse Response (IIR) filters (Same *et al.*, 2020) to process the coherence every T_{window}. The coherence γ^2 is averaged across the N sample frequencies from 200 Hz to 1250 Hz denoted f_{min} and f_{max} and converted to decibels to obtain the WIDs feature:

$$\Delta_{WID} = 10\log_{10}\left(\frac{1}{N}\sum_{f=f_{min}}^{f_{max}}\gamma^2(f)\right). \tag{2.2}$$

To minimize classification error, the resulting values are passed through a first-order unity-gain low-pass IIR filter with a cutoff frequency of 7.83 Hz. The filtered output is then compared against a threshold of -6 dB. Values below this threshold are considered WIDs.

If WIDs are detected, a VAD is used to confirm whether the participant is talking. The VAD extracts features from the IEM signal and makes decisions based on these features. For this

implementation, different features from the literature were assessed for their computational efficiency, information gain, and human interpretability. From this assessment, the selected features include the energy ratio between lower, mid, and high frequencies, $\frac{A_{f_{mid}}}{A_{f_{high}}}, \frac{A_{f_{mid}}}{A_{f_{low}}}$; the number of harmonics $N_{harmonics}$ in the mid-frequency range; and the evaluated SNR expressed in decibels and A-weighted decibels (Lezzoum *et al.*, 2014; Sohn *et al.*, 1999). The lower, mid, and higher frequencies denoted $f_{low}, f_{mid}, f_{high}$ range from 15 Hz to 153 Hz, 153 Hz to 1323 Hz, and 1323 Hz to 1944 Hz, respectively (Lezzoum *et al.*, 2014). Like the WIDsD, the VAD features are filtered through a first-order low-pass IIR filter.

The 5 extracted features are then compared to threshold values T_0, yielding binary logical outputs. These 5 outputs, denoted as $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_5)$, are input into a logistic regression model multiplied by gains $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_5)$, trained on a database that models the relationship between speech SPL, listener distance, and noise SPL (Bouserhal *et al.*, 2017a) to obtain the probability of a verbal event:

$$\Delta_{VAD} = \frac{1}{1 + e^{-\sum \beta_n \omega_n}} \mid 1 \leq n \leq 5. \tag{2.3}$$

Following the training, only probabilities of Δ_{VAD} above 50 % are considered to indicate an active talker.

If a participant is not identified as an active talker by the WIDsD nor the VAD and is not receiving speech from others, the signal is classified as noise. This noise signal is then used to update the noise SPL estimate through a first-order low-pass IIR filter.

2.4.4.4 Sound Spatialization

To spatialize sound, the OEM signal is filtered using a Talker-Listener Head Related Transfer Function (TL-HRTF). The TL-HRTF combines the voice directivity pattern and the HRTF. Initially, the relative orientations of the listener's and talker's heads are calculated. These relative orientations are then used with the TL-HRTF to determine the appropriate filter coefficients.

The absolute position and orientation of each participant are obtained from motion capture. Positions are on three axes: x, y, and z. For the n^{th} participant looking at the m^{th} participant, its rotation matrix $\mathbf{R}_n$ and their absolute Cartesian coordinates $\mathbf{C}_{a,n}, \mathbf{C}_{a,m}$ are used to calculate the relative Cartesian coordinates:

$$\mathbf{C}_{r,n\to m} = \mathbf{R}_n^{-1}(\mathbf{C}_{a,m} - \mathbf{C}_{a,n}). \tag{2.4}$$

These relative Cartesian coordinates are then converted into spherical coordinates:

$$\varphi_{n\to m} = \text{atan2}(\mathbf{C}_{ry,n\to m}, \mathbf{C}_{rx,n\to m}), \tag{2.5}$$

$$\theta_{n\to m} = \arctan\left(\frac{\mathbf{C}_{rz,n\to m}}{\sqrt{\mathbf{C}_{ry,n\to m}^2 + \mathbf{C}_{rx,n\to m}^2}}\right), \tag{2.6}$$

where $\varphi_{n\to m}$ and $\theta_{n\to m}$ are the relative azimuth and elevation angles, respectively.

The HRTF filters are sourced from the HUTUBS database and are based on the FABIAN head and torso simulator (Lindau, Hohn & Weinzierl, 2007; Brinkmann *et al.*, 2019). The voice directivity pattern is based on studies of the spherical directivity of the pressure field around the head (Dunn & Farnsworth, 1939). Each filter coefficient represents a gain and a phase shift applied to each FFT frequency. The TL-HRTF is implemented using three hash maps: two for the HRTF (one for each ear) and one for the voice directivity pattern. Each hash map stores filter coefficients that are applied to the audio for spatialization. The hash map stores 1536 amplitude and phase coefficients for each ear, for 440 positions. The selection of the appropriate filter is based on the relative azimuth $\varphi_{n\to m}$ and elevation $\theta_{n\to m}$. Once the correct filter is chosen, the voice directivity pattern filter is applied first, followed by the application of the HRTF filters in parallel to generate one signal for each ear.

2.4.4.5 Intended-Listener Detection

Once the talker is identified, the intended listeners must be determined with an ILD algorithm. The ILD is based on the concept of Communication Radius (CR). Every user within the CR is estimated to be an intended listener. The estimation is made with the Communication Radius Estimator (CRE). To enhance the distinction between listeners' CR, the mockups spatialize the signals before evaluating the CR. The CR is based on vocal effort. In this work, vocal effort is inferred from speech SPL L_W, noise SPL, and distance between users. The CR limit Δ_{CR} is defined by calculating a limit function based on the noise SPL $Noise_{SPL}$ and the logarithm of the distance d (Bouserhal *et al.*, 2017a):

$$\Delta_{CR} = 0.5 Noise_{SPL} + 0.8 \log_{10}(d) - 20. \tag{2.7}$$

Since speech is a dynamic signal that varies in intensity, this limit was adjusted downward to encompass most of the speech intensity distribution based on empirical tests. After estimating the Δ_{CR}, it is compared to the SPL of the spatialized OEM signal. If that SPL exceeds the Δ_{CR}, the participant is identified as an intended listener. In this work, all intended listeners receive the spatialized OEM signal.

2.4.4.6 Speech Enhancement

As previously mentioned, in this study, speech is recorded using the OEM, and noise is directly played into the participants' ears. This approach eliminates the need for noise reduction and bandwidth extension, while also providing better control over the noise SPL.

To regulate the speech SPL, an Automatic Gain Controller (AGC) is employed. The AGC boosts low-level speech and attenuates overly loud speech. It works by comparing the noise SPL with the speech SPL, then applying a gain to maintain a stable SNR of 10 dB in noisy scenarios and 30 dB in quiet scenarios.

2.4.4.7 Playing Audio

Once all digital signal processing is completed, the speech signal is converted back to the temporal domain using the Inverse Fast Fourier Transform (IFFT). To avoid discontinuities between speech processing loops, a windowing function based on the Hann window is applied. The Hann window length N is of 1024 points. The Hann window function multiplies each point n with a gain $w[n]$. The recorded spectral signal uses 10.6 ms from the previous loop and T_{window} from the current loop. The signal that is played must be of T_{window} long with minimal latency. To achieve this, the windowing function applies half of the Hann window to the first third of the IFFT signal, a gain of 1 to the second third, and the second half of the Hann window to the last third, as represented in Equation 2.8.

$$w[n] = \begin{cases} \frac{1-\cos\left(\frac{2\pi n}{N_{Hann}}\right)}{2} & \text{if } 0 \leq n < {}^{N}/_{3}, \\ 1 & \text{if } {}^{N}/_{3} \leq n < {}^{2N}/_{3}, \\ \frac{1+\cos\left(\frac{2\pi n}{N_{Hann}}\right)}{2} & \text{otherwise.} \end{cases} \tag{2.8}$$

Using the overlap-add method, the first two thirds of the windowed signal are combined with the last third of the previous loop's windowed signal. This windowing function results in a 32 ms latency at 48 kHz while ensuring continuity between speech processing loops. The threshold for perceiving latency is 68 ms in a factory noise environment and 26 ms in quiet conditions (Lezzoum *et al.*, 2016b). Therefore, the latency is expected to be slightly perceptible in quiet scenarios and imperceptible in noisy ones.

Finally, the windowed signal is combined with the noise and played through the intended listener's earpiece. The noise is sourced from a sound file, and its SPL is adjusted according to the scenario. In the quiet scenarios, the noise level is set to 55 dBA SPL white noise, while in the noisy scenario, it is set to 75 dBA SPL industrial noise.

2.4.5 Evaluation

The evaluation of the performance of RAVE as compared to the PTT is based on both subjective and objective data. Subjective data consists of participant survey responses and comments, offering qualitative insights into the user experience. Objective data encompasses all the recorded data during the experiment (including the audio and the position of participants) and the performance metrics of the various algorithms employed throughout the system.

2.4.5.1 Subjective Evaluation

After each scenario, participants completed the same set of 16 survey questions. These questions were related to each device's ease of use, including its usability at different distances (Questions 1 to 4) subsection 2.5.1, efficiency of communication (Questions 5 to 10) subsection 2.5.2, simulation anomaly (Questions 11 to 15) subsection 2.5.3, and an open-ended question for overall comments. The surveys questions are as follows:

Q1. How easy was it to use this device?

Q2-4. How much do you agree with this statement: It was easy for me to adjust my voice to speak with people at a long / medium / short distance?

Q5. How frequent did you notice you lost the beginning of what others were saying?

Q6. How much was lost in the beginning of what others were saying?

Q7. How often did communications that were not intended for you happen?

Q8. How much do you agree with this statement: Communication that were not intended for me bothered me?

Q9. How accurate was the perceived direction of the voice compared to the actual direction of the talkers?

Q10. How much do you agree with this statement: Having a directionality of voice helped my communication process?

Q11. How much do you agree with this statement: The transmitted voice was distinguishable from the original voice?

Q12. How much do you agree with this statement: There was a latency between what I hear and what I saw?

Q13. How much do you agree with this statement: The latency bothered me?

Q14. Which kind of artifact(s) (e.g., electrical noise, feedback) did you hear if any?

Q15. If any, how often did you hear the artifacts?

Q16. Do you have any other comments on the communication process you just tested?

Except for the open-ended questions, all the other questions used a 5-point Likert scale. In the descriptive statistics, mean responses were calculated by converting the ordered categorical scale to a numerical scale, i.e., 1, 2, 3, 4, and 5. The number of responses per ordered category was also reported. We further modelled each question's responses with ordinal mixed-effects regression using the ordinal package (Christensen, 2023) from R (R Core Team, 2022). The ordinal mixed-effects logistic regression models the probability that one condition receives a higher (or lower) rating on the Likert scale compared to another. The results are expressed as log-odds, which represent the natural logarithm of the odds of an outcome. To interpret the results, a log-odds of 0 means no difference between groups, a positive log-odds indicates greater odds of being rated in the higher category, and a negative log-odds indicates greater odds of being rated in the lower category. The magnitude represents the strength of this effect on a logarithmic scale. The modelling used variables representing the distribution, the type of device and the type of noise named `participant`, `device`, and `noise`, respectively. The variable `device` refers to either the PTT or the RAVE mockup. Similarly, `noise` is a variable referring to either the white or the factory noise. The random intercept of `participant` and the random slope given `device` was modelled. We tested the statistical significance of the main effects of `device` and `noise` and their interaction with ANOVA-based model fit comparison procedure. Post-hoc tests were conducted to determine the statistically significant effects.

2.4.5.2 Objective Evaluation

Using microphone recordings and participant position tracking, scenarios were reconstructed to evaluate the performance of the OVD and ILD algorithms. As detailed in subsubsection 2.4.4.3, the OVD comprises two components: the WIDsD and the VAD, both of which operate by comparing extracted features against predefined thresholds. Threshold performance was assessed using the TPR and False Positive Rate (FPR), which represent the proportions of correctly and falsely detected speech frames, respectively. The TPR should be maximized while the FPR minimized. Ground truth labels were manually generated based on the OEM recordings. Using this ground truth, the algorithms were optimized for both group-level and individual-level performance. For each participant, an optimal threshold set T_i was identified — defined as the threshold that maximizes the difference between TPR and FPR, thereby achieving the best possible trade-off between true and false detections. The performance of the ILD algorithm can be evaluated by verifying whether detected own-voice activity reached the intended listener—indicating a successful interaction. The head positions of the intended listeners can be confirmed using motion capture data, as participants were looking at one another while speaking. Similar to the evaluation of the OVD algorithm, the ILD algorithm can be trained individually for each participant under each condition for a personalized model and using the whole group for a non-personalized model. This involves modelling speech level as a function of background noise level and talker-to-listener distance, using a talker-dependent model (Bouserhal *et al.*, 2017a). Both personalized and non-personalized models can be trained from the collected data to define new Δ_{CR} models. For each interaction, the difference between the voice SPL (L_W) and the CR limit can be analyzed for the model used during the experiment, the non-personalized model, and the personalized models. A positive difference indicates that the intended listener was correctly identified, while a negative difference is a false negative. The greater this difference, the bigger the CR, which may also suggest that unintended listeners could have been targeted as well.

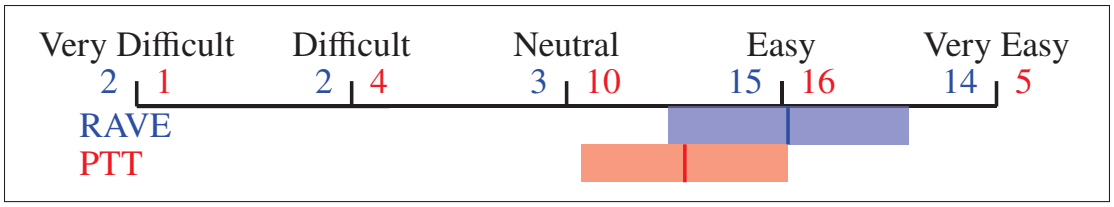

Figure 2.7 Answers to Q1: How easy was it to use this device?

2.5 Subjective Results

Results are presented with the raw data, statistical models, and comments from participants. For each question, a figure depicts the answers of the 18 participants. The numbers indicate the number of responses per ordered category per `device`. Both responses to noisy scenarios, i.e., white and factory are merged leading to 36 answers. The darker line represents the average score $\overline{x}$ per `device`. The transparent bar represents the data range $[\overline{x} - \sigma/2; \overline{x} + \sigma/2]$, where σ is the standard deviation.

2.5.1 Ease of Use

Figure 2.7 shows that, overall, while neither `device` was difficult to use, participants rated RAVE to be easier to use than the PTT. The average ease of use of the PTT mockup is half way between easy and neutral. The average ease of use of RAVE is slightly superior to easy. The main effect of `device` was statistically significant ($p < 0.05$). RAVE had a log-odds of 2.65 of being rated in a higher category in terms of ease of use than PTT; this difference was borderline statistically significant ($p = 0.065$). The main effect of `noise` was not statistically significant, neither was the interaction effect between `device` and `noise`.

Figure 2.8 shows that participants overall felt that adjusting their voices to reflect different communication distances was more difficult than using the PTT. At long distance, the main effect of `device` was statistically significant ($p < 0.05$). RAVE had a log-odds of 2.27 of being rated in a lower category in terms of ease of use than PTT ($p < 0.05$). The main effect of `noise` was not statistically significant, neither was the interaction effect between `device` and `noise`.

When the distance is medium or short, the interaction effect between `noise` and `device` was statistically significant (Medium: $p < 0.05$, Short: $p < 0.01$). When the distance is medium, the main effect of `device` is statistically significant ($p < 0.05$), and the main effect of `noise` has borderline statistical significance ($p = 0.069$). RAVE had a log-odds of 3.47 of being rated in a lower category than PTT ($p < 0.05$). The white-noise condition had a log-odds of 1.79 of being rated in a lower category than in factory noise ($p < 0.05$). However, RAVE in white noise had a log-odds of 2.43 to be rated in a higher category ($p < 0.05$). When the distance is short, the main effects of `device` and `noise` are both statistically significant ($p < 0.01$). RAVE had a log-odds of 3.64 of being rated in a lower category than PTT ($p < 0.05$), the white-noise condition had a log-odds of 2.51 of being rated in a lower category than in factory noise ($p < 0.01$), yet RAVE in white noise had a log-odds of 3.13 to be rated in a higher category ($p < 0.05$). On the factory noise RAVE scenarios, a participant noted that the speech level required was very uncomfortable. On PTT scenarios, some participants struggled familiarizing themselves with the timing of pressing and releasing the button. A participant noted that without visual feedback, it was hard to know if the intended listeners were really targeted, regardless of the scenario.

2.5.2 Communication Efficiency

As shown in Figure 2.9, participants had the impression of losing a little bit less than a word of what others were saying a little bit more than occasionally for PTT and RAVE scenarios. Concerning the PTT scenarios, some participants found that the button had some latency. In the RAVE scenarios, some participants attempted to first get the attention of the intended listeners by calling their name before speaking, but not all participants followed this approach. In terms of the frequency of lost beginnings, neither the main effect of `noise` or `device` nor their interaction was statistically significant. In terms of the amount, both the white-noise condition and RAVE condition had higher odds ratio to be rated in a lower category, but neither effect was statistically significant. The interaction effect reached a borderline statistical significance

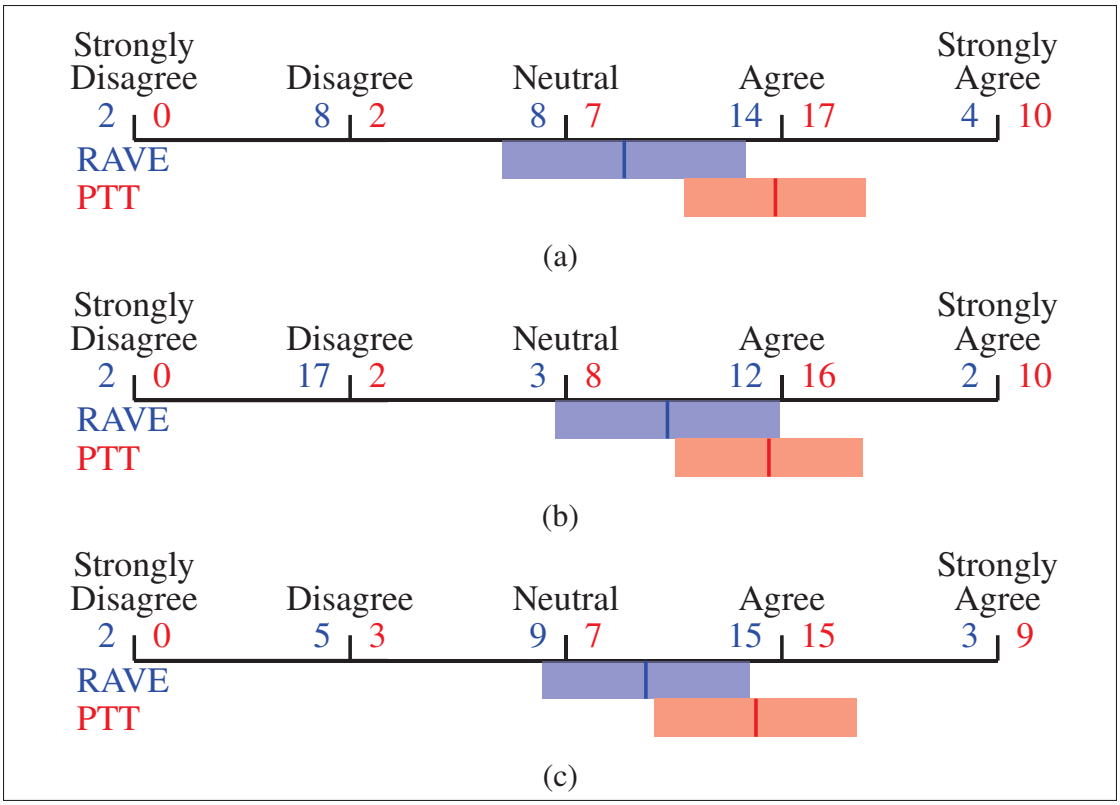

Figure 2.8 Answers to Q2/3/4: How much do you agree with this statement: It was easy for me to adjust my voice to speak with people at a long/medium/short distance? (a)/(b)/(c)

($p = 0.065$); RAVE in white noise was rated with a 1.79 odds ratio to be in a higher category ($p = 0.072$).

As shown in Figure 2.10, both `device` conditions have non-intended listeners contacted a little bit more than often, with RAVE being a little bit better. Participants reported being neutral regarding how bothersome it was. In terms of the frequency of non-intended communication, while RAVE had a log-odds of 0.62 of being rated in the higher category, this findings was not statistically significant, and neither was the main effect of `noise` nor their interaction. In terms of the level of annoyance, there was no statistically significant effect for `noise` or `device`.

As shown in Figure 2.11, participants reported an impression of a direction with a precision of around 25 degrees with RAVE scenarios and around 40 degrees with PTT scenarios. Some

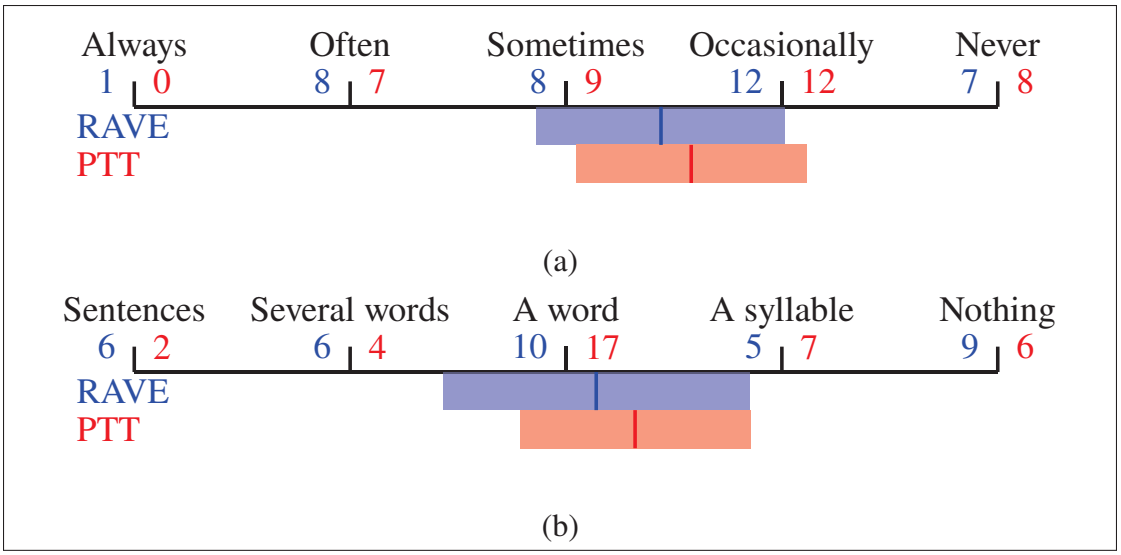

Figure 2.9 (a) Answers to Q5: How frequent did you notice you lost the beginning of what others were saying? (b) Answers to Q6: How much was lost in the beginning of what others were saying?

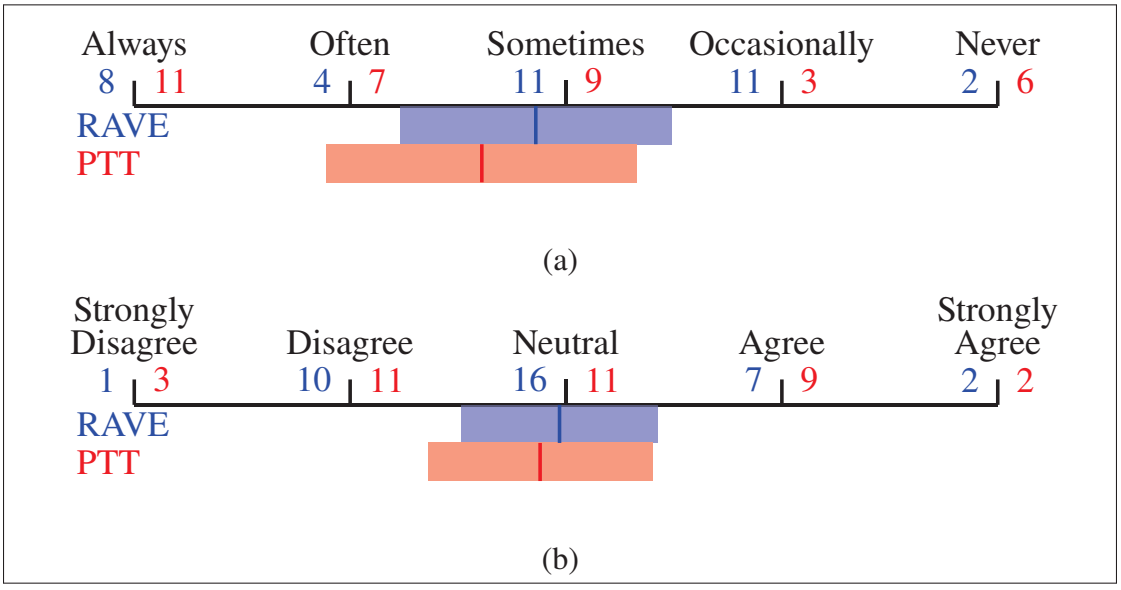

Figure 2.10 (a) Answers to Q7: How often did communications that were not intended for you happen? (b) Answers to Q8: How much do you agree with this statement: Communication that were not intended for me bothered me?

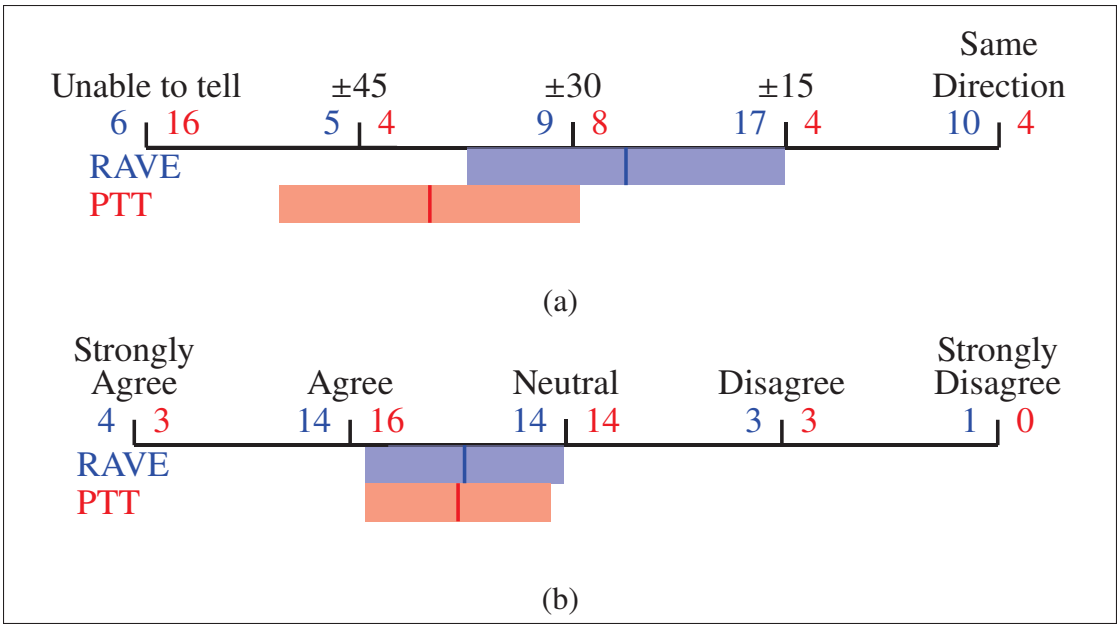

Figure 2.11 (a) Answers to Q9: How accurate was the perceived direction of the voice compared to the actual direction of the talkers? (b) Answers to Q10: How much do you agree with this statement: Having a directionality of voice helped my communication process?

participant verbally expressed confusion between their front and back. One participant noted a confusion between the left and right. No effect from `device` nor `noise` was found for Question 10. Participants felt that having a directionality slightly improved communication. In terms of the accuracy of directionality, statistical significance was reached for the main effects of `noise` ($p < 0.05$) and `device` ($p < 0.01$) and their interaction effect ($p < 0.05$). The white-noise condition had a log-odds of 1.26 of being rated in a higher category ($p < 0.1$) and the RAVE condition had an odd ratio of 3.68 of being rated in a higher category ($p < 0.01$); although, there is an interaction effect that the white noise-RAVE condition had a 2.62 odds ratio of being rated in a lower category ($p < 0.05$). No statistically significant difference was seen when asked how the directionality helped the communication.

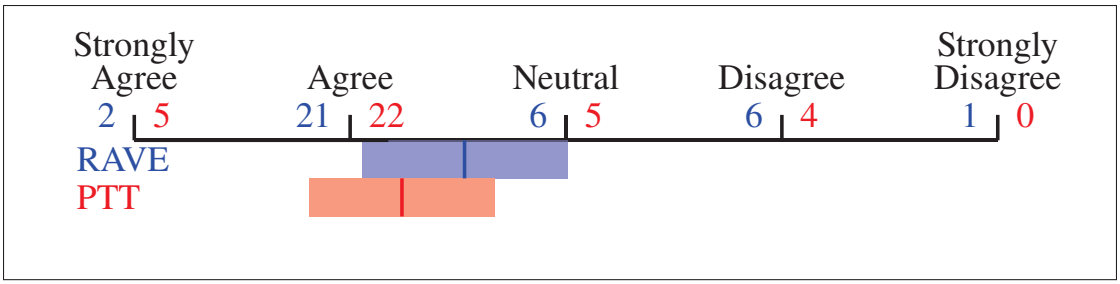

Figure 2.12 Answers to Q11: How much do you agree with this statement: The transmitted voice was distinguishable from the original voice?

2.5.3 Mockup Realism

Figure 2.12 shows that participants reported that they could differentiate the transmitted voice from the voice from the passive path. No statistically significant difference was found between `device` or `noise`. Figure 2.13 shows that participants in general did not experience much latency, and that latency did not disturb the communication process. No statistically significant difference was found between `device` or `noise`. These interferences are labelled as noise artifacts. No statistically significant difference was found between `device` or `noise`. The responses from the participants varied a lot, with “A couple of times" as the most reported answer as seen in Figure 2.14. There was a borderline statistically significant main effect of sound condition ($p = 0.082$), where the white noise condition had a log-odds of 0.84 of being in a higher category ($p = 0.086$); otherwise, there was no statistically significant difference found between conditions.

2.6 Objective Results

2.6.1 Own-Voice-Detection Algorithm Performance

Different thresholds give a different TPR and FPR as represented for the WIDsD in Figure 2.15. The mean ideal thresholds, denoted as $\overline{T_i}$, for all features are shown in Table 2.1. The actual mean performance between participants with the threshold T_0 are shown in Table 2.1 column 4

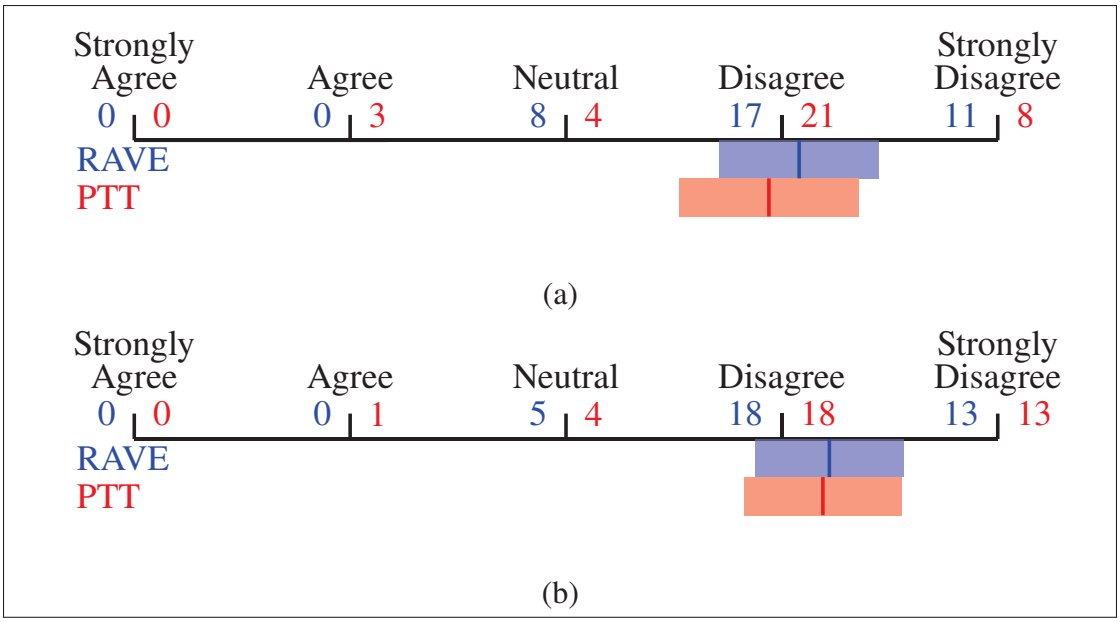

Figure 2.13 (a) Answers to Q12: How much do you agree with this statement: There was a latency between what I hear and what I saw? (b) Answers to Q13: How much do you agree with this statement: The latency bothered me?

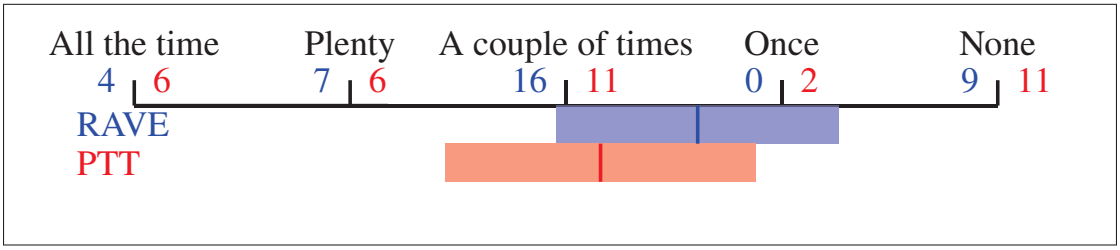

Figure 2.14 Answers to Q15: If any, how often did you hear the artifacts?

and 7. The mean performance with the mean ideal thresholds $\overline{T_i}$ and the ideal thresholds T_i are also shown in columns 5, 6, 8 and 9. For most features, ideal thresholds lead to just a slight improvement compared to the use of the mean average thresholds.

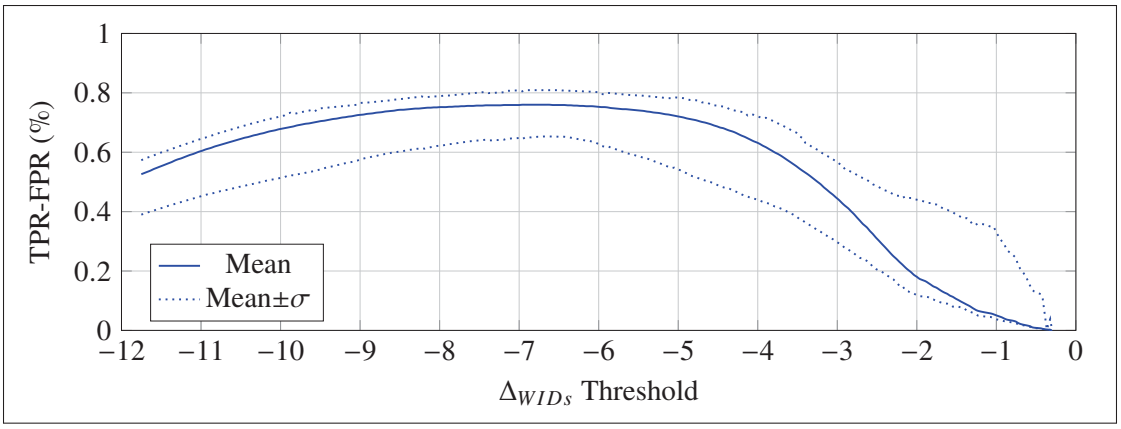

Figure 2.15 Performance criterion in function of the threshold used for the WIDsD

Table 2.1 Comparison of the average performance between the thresholds used during data collection T_0, the mean ideal thresholds $\overline{T_i}$, and the ideal thresholds T_i in terms of TPR and FPR

	Threshold		TPR (%)			FPR (%)		
Features	T_0	$\overline{T_i}$	T_0	$\overline{T_i}$	T_i	T_0	$\overline{T_i}$	T_i
WIDsD	6.00	6.69	85.8	91.3	91.0	13.1	16.0	10.8
$\frac{A_{f_{mid}}}{A_{f_{low}}}$	1.67	0.39	67.2	82.8	78.1	36.5	47.8	23.8
$\frac{A_{f_{mid}}}{A_{f_{high}}}$	67.36	53.3	63.1	69.3	72.9	16.7	19.7	16.0
$N_{harmonics}$	0.86	1.97	90.2	65.8	73.9	57.1	19.2	19.2
SNR dB	4.40	4.70	72.7	70.3	77.4	45.4	40.9	38.3
SNR dBA	4.77	3.96	68.1	73.4	79.1	39.6	43.5	42.5
Δ_{VAD}	0.50	0.29	78.2	81.0	84.6	4.7	6.1	6.5

2.6.2 Intended-Listener-Detection Algorithm Performance

For each interaction, the distance between the talker and the intended listener as a function of the difference between the voice SPL (L_W) and the CR limit (Δ_{CR}) is shown in Figure 2.16. The average results of the different models are summarized in Table 2.2. For each model and across various distance ranges, the mean differences between L_W and Δ_{CR}, along with the standard deviations, are reported. As shown in Table 2.2, using a model personalized results in a mean

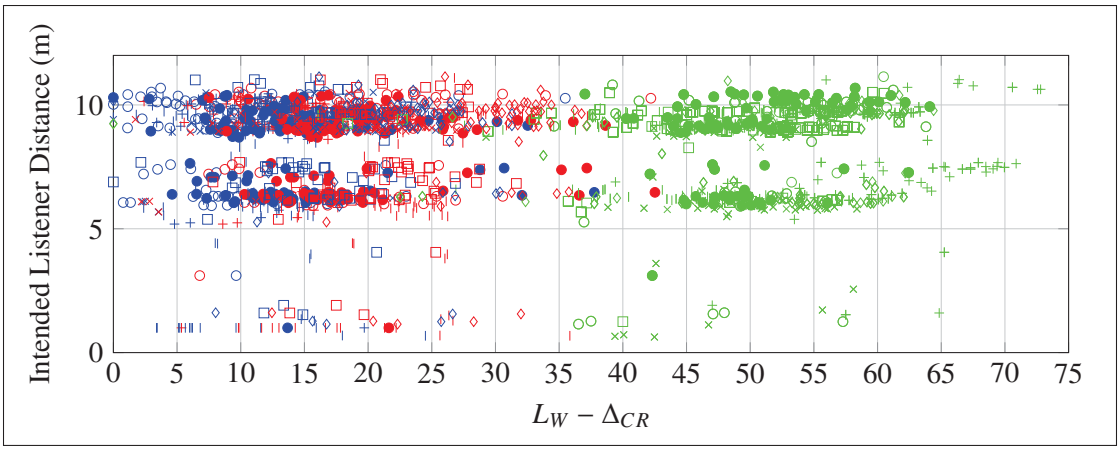

Figure 2.16 Intended listener distance as a function of the difference between the IEM speech SPL (L_W) and the CR threshold (Δ_{CR})

The IEM speech SPL (L_W) and the CR threshold (Δ_{CR}) is shown for the personalized model (blue), the non-personalized model (red), and the model used during data collection (green). Different marker styles denote individual participants

Table 2.2 Comparison of the average performance between the model used during data collection Δ_{CR_0}, the non-personalized model $\overline{\Delta}_{CR_i}$, and the personalized model Δ_{CR_i} for different distances

Model used	Δ_{CR_0}	$\overline{\Delta}_{CR_i}$	Δ_{CR_i}
mean (Lw - Δ_{CR}) between 8–12 m	50.4	23.5	15.5
std.dev (Lw - Δ_{CR}) between 8–12 m	8.1	8.0	7.3
mean (Lw - Δ_{CR}) between 5–8 m	51.3	25.2	15.3
std.dev (Lw - Δ_{CR}) between 5–8 m	8.0	7.8	7.0
mean (Lw - Δ_{CR}) between 0–5 m	49.1	22.2	13.4
std.dev (Lw - Δ_{CR}) between 0–5 m	7.1	7.4	6.0

value closer to zero and a smaller standard deviation. A mean closer to zero indicates higher accuracy, while a smaller standard deviation reflects greater precision of the algorithm.

2.7 Discussion

The advantages and limitations of RAVE can be assessed based on the objective and subjective results. The assessment needs to be taken into context, as the data collection was done in a

limited controlled environment with untrained participants by comparing PTT with RAVE. Since RAVE is based on communication functionalities such as OVD, ILD, and TL-HRTF, this section discusses the accuracy of our approach with regard to these functionalities.

2.7.1 Overall Realism

Overall, the data collection setup was realistic for small groups of users with normal hearing as shown in the survey. As reflected in questions 12 and 13 (Figure 2.13), the latency in the mockup implementation was barely noticeable to participants and did not affect their experience. The 32 ms latency fell within acceptable limits established by previous research (Lezzoum *et al.*, 2016b), ensuring minimal perceptual impact. Questions 14 and 15 (Figure 2.14) show that participants noticed some artifacts during the data collection, but its impact was limited. Artifacts were more noticeable in quiet conditions, whereas industrial noise may have masked them in noisier environments. Regardless, even with the presence of artifacts, participants were able to complete their tasks, suggesting that the artifacts were not too disruptive. Therefore, the experiment was, to a certain degree, immersive enough for participants to test the technology.

2.7.2 Detecting Own Voice Activity

Regarding the OVD which has been used to send speech automatically, its implementation in this work was satisfactory. As shown by the answers to questions 5 and 6 (Figure 2.9), the OVD performance is comparable to using a PTT system, as the effect of the device on the amount and the frequency of words lost was not statistically significant. With similar performances to the PTT, the OVD is still perceived as beneficial, as reflected in survey question 1 (Figure 2.7). It is probably beneficial as it eliminates the need for participants to manually press a button. Similar to other studies on OVD (Pertilä *et al.*, 2021), personalizing the algorithm for individual users does not lead to significant overall performance improvements, as shown in Table 2.1. The TPR and FPR do not change significantly between the uses of the mean ideal thresholds and the ideal personalized thresholds. However, personalizing some specific parameters—such as the

mid-to-low frequency ratio and the number of harmonics—substantially impacts performance and should therefore be considered.

2.7.3 Detecting Intended Listener

As mentioned in subsubsection 2.4.4.5, the implemented ILD was based on a CR. While TPR for this feature was perfect, survey questions 2, 3, and 4 (Figure 2.8), along with participant feedback, indicated that the ILD was less intuitive than expected. A potential improvement would be to provide visual feedback, e.g., through LEDs, to indicate when other users are receiving speech. As shown, personalized models to determine Δ_{CR} offer better performance as they reduce variance and do not need large limit margins. However, using talker-dependent mode, the ILD algorithm could be improved.

While the ILD is expected to prevent disturbance to non-intended listeners, this effect was not observable in the collected data. As shown with survey questions 7 and 8 (Figure 2.10), the participants were not bothered by unintended communication which can be explained by the small number of participant in each group. A larger participant group would likely have increased the number of unintended communications, thereby amplifying the observable effects of RAVE and PTT. Nonetheless, ILD shows potential as it allows to dynamically select intended listeners.

2.7.4 Spatializing Sound

Sound spatialization was achieved with the TL-HRTF as described in subsubsection 2.4.4.4. As shown with survey question 9 (Figure 2.11), the TL-HRTF have a positive impact on the perceived direction of speech. Some participants still experienced front and back or left and right confusion. Given the individual variability in TL-HRTF performance, implementing a personalized model for each user would help reduce perceptual confusion. Some work tends to link personalized HRTF and morphological features (Pelzer *et al.*, 2020).

Similar to the case with ILD, participants did not perceive a noticeable benefit from incorporating directionality into the communication. As data collection was conducted exclusively with groups of three participants, larger group sizes would likely have amplified the impact of directional communication.

2.8 Conclusion

This study addressed gaps in previous research by integrating and linking components of the proposed Radio Acoustical Virtual Environment (RAVE) technological approach, which had previously been studied in isolation. The development of a RAVE mockup enabled this study to present participant feedback and performance evaluations for key algorithms, including the own voice detection algorithm, the intended listener detection algorithm, and the talker-listener head-related transfer function algorithm. The findings demonstrate potential benefits in enabling users to dynamically address specific individuals through vocal effort and transforming speech signals to convey spatial directionality. This work provides insights into how user interaction with Active Hearing Protection Device (AHPD) communications systems and highlight the importance of personalized algorithms. The results will inform future development of AHPD communications systems.

2.9 Acknowledgments

The authors would like to acknowledge the financial support received from NSERC Alliance (ALLRP 566678-2021), MITACS IT26677 (SUBV-2021-168) and PROMPT (#164_Voix-EERS 2021.06), for the ÉTS-EERS Industrial research chair in in ear technologies, sponsored by EERS Global Technologies. The support provided by CIRMMT for the facility, as well as the technical assistance from Yves Méthot and the entire team of CIRMMT technicians, is deeply appreciated.

CONCLUSION ET RECOMMANDATIONS

Ce mémoire présente le travail accompli dans le cadre du programme de maîtrise portant sur la valeur ajoutée de la technologie émergente d'environnement radio-acoustique virtuel (*radio acoustical virtual environment*) (RAVE). Cette section offre une synthèse du travail réalisé ainsi que des recommandations pour de futurs travaux.

3.1 Synthèse

Trois objectifs secondaires ont été atteints afin de réaliser l'objectif principal de ce mémoire : mesurer la valeur ajoutée de RAVE.

Premièrement, un banc d'essai pour RAVE a été implémenté, tel que décrit dans la sous-section 2.4.2, et testé avec des participants, comme discuté dans la sous-section 2.4.4. Cet objectif a été atteint en adaptant plusieurs algorithmes existants pour une utilisation en temps réel, notamment la détection de la voix de l'utilisateur (*Own Voice Detection*) (OVD), la détection de l'auditeur visé (*Intended Listener Detection*) (ILD) et la fonction de transfert liée aux têtes de l'auditeur et du locuteur (*Talker-Listener Head Related Transfer Function*) (TL-HRTF). L'OVD repose sur la lecture de différents descripteurs interprétables et prend une décision fondée sur leurs états. L'ILD, quant à lui, s'appuie sur le concept d'estimateur du rayon de communication (*Communication Radius Estimator*) (CRE). Ce dernier renvoi un rayon de communication (*Communication Radius*) (CR) en fonction de l'effort vocal estimé. Un effort vocal plus important donne un plus grand CR. Les utilisateurs situés plus près du locuteur que le CR sont considérés comme les auditeurs visés. La TL-HRTF permet de spatialiser le son en appliquant des filtres sélectionnés selon l'orientation des participants. Certaines simplifications ont été nécessaires, et le banc d'essai n'intègre pas encore d'algorithmes de réduction de bruit ni d'extension de bande-passante (*Bandwidth Extansion*) (BWE). Néanmoins, il s'est avéré suffisamment réaliste pour permettre une évaluation pertinente de la technologie RAVE dans des

scénarios en temps réel. Les essais réalisés avec des participants ont démontré que le banc d'essai était perçu comme acceptable. La majorité des utilisateurs n'a pas remarqué de latence notable dans le traitement du signal. Quelques artefacts ont été relevés, mais sans impact significatif sur l'expérience globale.

Le deuxième objectif secondaire de ce travail consistait à créer une base de données incluant la position en temps réel des participants, ainsi que des enregistrements de leur voix — bruitée et non bruitée — à l'aide de microphones extra-auriculaire (*Outer-Ear Microphones*) (OEMs) et de microphones intra-auriculaire (*In-Ear Microphones*) (IEMs). Pour atteindre cet objectif, le banc d'essai de RAVE ainsi qu'un banc d'essai équipé d'un système peser pour parler (*Push-To-Talk*) (PTT) ont été testés par des participants. Ces essais ont permis de recueillir les données nécessaires. En plus de nourrir les conclusions de ce mémoire, ces données pourront servir de base à de futurs travaux sur RAVE ou sur d'autres types de protecteurs auditifs actifs (*Active Hearing Protection Devices*) (AHPDs). Elles ont été enregistrées dans la banque de données nommée CRITIAS :DB.

Enfin, le troisième objectif secondaire visait à identifier des pistes d'amélioration pour de futures itérations de RAVE, ainsi que pour les systèmes de communication intégrés aux AHPDs. Les tests effectués avec des participants ont permis de recueillir à la fois des données subjectives (avis des participants) et objectives (performances réelles des algorithmes), comme discuté dans la section 2.5 et la section 2.6, respectivement. L'algorithme d'OVD s'est montré fonctionnel, bien que son implémentation puisse être optimisée. Les utilisateurs ont exprimé un avis favorable quant à sa présence, et il est important de souligner qu'il n'est pas nécessaire de le personnaliser pour chaque individu. En revanche, les algorithmes d'ILD et de TL-HRTF bénéficieraient à être personnalisés afin d'en améliorer les performances. Plus spécifiquement, l'algorithme d'ILD devrait être repensé pour gagner en précision, tandis que la TL-HRTF, bien que jugée efficace

par la majorité des participants, pourrait être ajustée individuellement afin de limiter les erreurs de localisation chez certains utilisateurs.

Ces trois objectifs secondaires ont permis de répondre à l'objectif principal du mémoire. Tel que résumé dans la section 2.8, la valeur ajoutée de RAVE réside dans ses deux fonctionnalités principales :

- La capacité à transmettre dynamiquement la voix d'un utilisateur à un autre utilisateur ciblé selon l'effort vocal, grâce à l'OVD et à l'ILD ;
- La spatialisation du son permettant de donner une direction perceptible à l'aide de la TL-HRTF.

L'OVD permet de libérer les mains des utilisateurs, ce qui facilite l'utilisation. Bien que fonctionnel, l'ILD n'a pas été aussi bien perçu. Outre le fait que le concept de CRE n'est peut-être pas optimal, le manque de rétroaction rendait l'utilisation de l'ILD moins intuitive. Un simple voyant lumineux indiquant les auditeurs ciblés pourrait résoudre ce problème et rendre RAVE beaucoup plus intuitif. La TL-HRTF a été perçue comme légèrement bénéfique. Probablement que la spatialisation du son aurait un plus grand potentiel avec un plus grand groupe d'utilisateurs. Ainsi, les fonctionnalités de RAVE se présentent vraiment comme le futur des systèmes de communication avec des AHPDs. Cependant, l'implémentation de ces fonctionnalités devra s'inscrire dans une continuité de travaux visant à concrétiser pleinement la vision de RAVE.

3.2 Recommandations

Ce mémoire s'inscrit dans la continuité d'une série de projets visant à améliorer la communication en milieu industriel. Bien que les objectifs définis aient été atteints, des travaux supplémentaires seront nécessaires pour poursuivre cette amélioration. Dans une démarche itérative et réaliste, des progrès significatifs peuvent être réalisés. Cette section présente ainsi plusieurs recommandations pour les prochaines phases du projet.

3.2.1 Améliorer l'algorithme ILD

La technologie RAVE constitue une évolution logique par rapport à l'utilisation d'un simple OVD. En plus de détecter les moments où l'utilisateur parle, elle introduit la possibilité d'identifier l'auditeur visé, ce qui pourrait rendre les communications bien plus dynamiques dans des environnements bruyants. Toutefois, le concept de rayon de communication fondé sur l'effort vocal n'est pas aussi intuitif qu'il pourrait paraître.

En effet, bien que les locuteurs ajustent naturellement leur effort vocal en fonction du niveau de bruit ambiant, de l'occlusion entre interlocuteurs, et de la distance avec la personne visée, la relation entre ces facteurs reste complexe et difficile à exploiter de manière fiable. Pour rendre le système plus efficace, il serait idéal d'intégrer un mécanisme de rétroaction permettant au locuteur de confirmer les auditeurs réellement visés.

Par ailleurs, un algorithme dédié devrait être développé pour identifier les auditeurs ciblés avec une plus grande précision. Les modèles existants, fondés sur les relations décrites par Bouserhal *et al.* (2017a), Traunmüller & Eriksson (2000) et Pelegrín-García *et al.* (2011), bien qu'intéressants, ne semblent pas suffisants à eux seuls pour couvrir toute la complexité du phénomène.

3.2.2 Intégration d'un module d'extension de bande et de réduction de bruit dans le banc d'essai

Les efforts de développement du banc d'essai présenté dans ce mémoire se sont principalement concentrés sur l'implémentation dans un système fonctionnel cohérent de trois modules clés : un algorithme d'OVD, un algorithme d'ILD, et un module de TL-HRTF.

Pour renforcer la robustesse et le réalisme du banc d'essai, il serait pertinent d'y intégrer deux modules supplémentaires : un module de BWE et un module de réduction de bruit. L'ajout de

ces modules permettrait d'utiliser le banc d'essai dans des environnements réels bruyants, plutôt que de simuler le bruit ambiant en injectant directement des signaux sonores dans les bouchons des participants. Une telle approche rendrait l'évaluation de la technologie plus représentative des conditions d'utilisation sur le terrain.

Des algorithmes pertinents ont déjà été proposés dans la littérature, notamment par Bouserhal *et al.* (2017b) et Lezzoum *et al.* (2016a). Bien qu'ils puissent nécessiter certaines adaptations pour une exécution en temps réel, ces ajustements semblent techniquement réalisables.

3.2.3 Tester le banc d'essais avec des utilisateurs ayant des problèmes auditifs

Comme l'ont démontré plusieurs études, le port de protecteurs auditifs (*Hearing Protection Devices*) (HPDs), combiné à un environnement bruyant, nuit encore davantage à l'intelligibilité de la parole chez les personnes ayant des troubles de l'audition (Lindeman, 1976; Abel *et al.*, 1982; Giguère & Dajani, 2009). Or, ce sont précisément ces utilisateurs qui pourraient le plus bénéficier d'un système tel que RAVE.

Il serait donc essentiel de tester le banc d'essai auprès de cette population afin de vérifier que la technologie est fonctionnelle pour des individus présentant divers profils auditifs. Leurs commentaires pourraient différer de ceux d'utilisateurs sans déficience auditive, et ainsi contribuer à dresser un portrait plus complet des besoins et attentes vis-à-vis de la technologie RAVE.

3.2.4 Modéliser le dérangement du délai de transmission

L'excellent travail de Lezzoum *et al.* (2016b) propose une modélisation du seuil de perception du délai de transmission. Toutefois, cette étude mériterait d'être approfondie afin de mieux comprendre les interactions entre le niveau de bruit ambiant, le délai de transmission et le degré de gêne ressentie par l'utilisateur.

Établir une telle relation permettrait de mieux évaluer les compromis possibles entre la longueur du délai et le niveau de dérangement perçu. Ce compromis pourrait s'avérer déterminant pour guider certaines décisions d'implémentation, notamment en ce qui concerne l'architecture des modules de RAVE ou le choix des technologies de transmission audio à privilégier.

3.2.5 Développement d'un matériel pour des tests sur le terrain

Le banc d'essai a été développé à l'aide d'équipements relativement encombrants, ce qui en limite la portabilité. Cette contrainte réduit la mobilité du système, obligeant son utilisation dans un environnement de laboratoire dédié, nécessitant souvent une réservation préalable. Les principaux équipements à l'origine de cette limitation sont les radios VHF utilisées pour la transmission audio, ainsi que les caméras de capture de mouvement.

Pour améliorer la portabilité du système, il serait pertinent d'envisager le remplacement d'un arrangement avec des radios VHF par un arrangement de transmission plus compacte, telles qu'un arrangement utilisant le Wi-Fi, le Bluetooth ou encore l'utilisation de radios ultras large bande. Le Wi-Fi et le Bluetooth bien que largement disponible peuvent introduire un délai de transmission plus important, qui doit rester acceptable d'un point de vue perceptif. Une meilleure compréhension de la relation entre ce délai et le niveau de gêne, comme mentionné précédemment, serait essentielle pour évaluer la faisabilité de ce compromis. Sinon, l'utilisation de radios ultra large bande, notamment celles de Spark Microsystems offre une très faible latence, une très faible consommation d'énergie et un débit amplement suffisant pour RAVE.

Concernant la capture de mouvement, bien que les caméras utilisées offrent une précision au millimètre près et ne présentent pas de dérive de position, leur déploiement est limité à une zone d'environ 10 mètres et nécessite un espace équipé spécifiquement. Il serait donc souhaitable d'explorer des solutions alternatives, moins encombrantes et plus flexibles quant à la position des utilisateurs, même si cela implique une légère diminution de la précision.

3.2.6 Inclusion d'un module de contrôle actif du bruit

Le rôle principal des AHPDs est d'assurer une protection efficace contre les effets nocifs du bruit, en offrant une bonne atténuation, souvent mesurée par la cote de réduction du bruit (*Noise Reduction Rating*) (NRR). L'une des techniques couramment employées pour améliorer ce paramètre est l'intégration d'un système de contrôle actif du bruit (*Active Noise Control*) (ANC).

Or, la version de RAVE décrite par Bou Serhal *et al.* (2013) n'intègre pas explicitement de fonctionnalité d'ANC. Il serait donc pertinent d'explorer l'ajout d'un tel système, notamment en exploitant la configuration de capture vocale basée sur l'agencement des IEMs, telle que proposée dans le présent mémoire. Une telle intégration pourrait non seulement améliorer la protection auditive, mais aussi renforcer la qualité de la capture vocale dans des environnements bruyants.

3.2.7 Vérifier la conscience spatiale avec un module de contrôle actif du bruit

La capacité à localiser la direction des sons dans les environnements industriels est essentielle pour des raisons de sécurité, comme la détection d'une alarme de recul ou d'autres signaux critiques. L'intégration d'un module d'ANC ainsi que les autres composantes de RAVE, devrait permettre à l'utilisateur de percevoir et localiser ce type de signaux sonores. Il est toutefois nécessaire d'évaluer dans quelle mesure ces composantes impactent la conscience spatiale d'un utilisateur. Une attention particulière devra être portée à la préservation des indices spatiaux tout au long du traitement audio, afin de garantir que ces signaux demeurent perceptibles et directionnels.

ANNEXE I

FORMULAIRE D'INFORMATION ET DE CONSENTEMENT

Les documents suivants contiennent les formulaires d'information et de consentement en français et en anglais. Tous les participants ont lu et signé le formulaire d'information et de consentement.

RÉSERVÉ au CÉR

Ne rien écrire dans cette case

Gabarit FIC BD - version 14 avril 2020

FORMULAIRE D'INFORMATION ET DE CONSENTEMENT

TITRE DU PROJET DE RECHERCHE

Analyse de la valeur ajoutée d'un système de communication dans un environnement radio acoustique virtuelle (RAVE)

CHERCHEURE RESPONSABLE

Pascal Giard, Professeur – École de technologie supérieure, Département de génie électrique

CO-CHERCHEURS ET COLLABORATEURS

Rachel Bouserhal, Professeure au département de génie électrique – École de technologie supérieure (ÉTS)

Jérémie Voix, Professeure au département de génie mécanique – École de technologie supérieure (ÉTS)

Xinyi Zhang, étudiante au doctorat au département de génie électrique – École de technologie supérieure (ÉTS)

Yves Méthot coordonnateur de l'électronique – CIRMMT

Julien Boissinot gestionnaire des systèmes/directeur technique – CIRMMT

Sylvain Pohu Responsable de production – CIRMMT

Valentin Pintat, Responsable de laboratoire à CRITIAS – École de technologie supérieure (ÉTS)

Dany Morissette, Assistant de recherche– École de technologie supérieure (ÉTS)

Arianne Marcoux, Assistante de recherche– École de technologie supérieure (ÉTS)

ÉTUDIANTE

Gabriel Ouaknine-Beaulieu, Étudiant à la maitrise au département de génie électrique – ÉTS

FINANCEMENT

Pascal Giard a reçu du financement du Conseil de recherches en sciences naturelles et en génie du Canada dans le cadre de la chaire CRITIAS.

INTRODUCTION

Nous vous invitons à participer à un projet de recherche. Cependant, avant d'accepter de participer à ce projet et de signer ce formulaire d'information et de consentement, veuillez prendre le temps de lire, de comprendre et de considérer attentivement les renseignements qui suivent.

Ce formulaire peut contenir des mots que vous ne comprenez pas. Nous vous invitons à poser toutes les questions que vous jugerez utiles à la chercheure responsable de ce projet ou à un membre de l'équipe de recherche, et à leur demander de vous expliquer tout mot ou renseignement qui n'est pas clair.

NATURE ET OBJECTIFS DU PROJET DE RECHERCHE

L'objectif de ce projet de recherche est d'évaluer la valeur ajoutée d'un environnement virtuel acoustique radio dans un système de communication.

Le projet de recherche vise également à contribuer à une banque de données établie à des fins de recherche qui est intitulée CRITIAS : DB. Le but de cette banque de données est de regrouper toutes les données collectées par divers projets menés par le Professeur Jérémie Voix et Rachel Bousheral et de les mettre à la disposition des chercheurs de l'École de technologie supérieure et d'autres institutions, dans le but de faire progresser les connaissances dans le domaine de la santé et de la sécurité au travail.

Pour mener à bien le projet, nous avons l'intention de recruter 30 participants, de tous genres, âgés de 18 ans ou plus.

DÉROULEMENT DU PROJET DE RECHERCHE

I. Lieu de réalisation du projet de recherche et durée de la participation

Une présélection aura lieu à la Chaire de recherche industrielle ÉTS-EERS en technologies intra-auriculaires (CRITIAS) au 355 Peel, Montréal. La collecte de données se déroulera à l'Université McGill au Centre de recherche interdisciplinaire en médias et technologies musicales (CIRMMT) au 527 Sherbrooke Ouest, Montréal. Votre participation totale au projet durera 1 heure et 30 minutes et nécessitera 2 visites.

II. Nature de votre participation

1. Évaluation de l'admissibilité et collecte d'informations démographiques (environ 20 mins)

Tout d'abord, votre éligibilité à la participation sera évaluée à CRITIAS. Vous remplirez un formulaire d'éligibilité et effectuerez une inspection otoscopique. L'inspection otoscopique évalue l'état de votre oreille externe et de votre tympan. Ensuite, vous passerez un test audiométrique tonal. Ce test vise à évaluer la sensibilité de votre ouïe. Pendant ce test, vous entendrez différents sons à différentes intensités et fréquences. Vous devrez répondre sur une tablette si vous pouvez entendre le son ou non. À la fin de cette étape, il est possible que vous ne puissiez pas passer aux étapes suivantes. Si tel est le cas, le chercheur expliquera les raisons. Si vous êtes éligible, la largeur de votre tête, l'espace entre vos oreilles et votre visage, ainsi que quelques autres mesures de votre visage, seront prises pour sélectionner des filtres audio qui vous seront le plus individualisés possible à l'aide d'un algorithme d'apprentissage machine.

	
Fig. 1 Otoscope et bouchon	Fig. 2 Kit de test audiométrique tonal

2. Collecte de données (55 mins)

Vous serez jumelé à deux autres participants et vous évoluerez ensemble dans quatre scénarios de dix minutes. Dans chaque scénario, vous porterez un appareil de communication et un casque de sécurité avec un dispositif de suivi. Ces appareils vous permettront de communiquer avec les autres participants et de créer l'effet d'environnement virtuel. Vous porterez également une ceinture qui enregistre votre rythme cardiaque. Cette ceinture améliorera la base de données d'une autre recherche. La moitié des scénarios seront avec l'effet d'environnement virtuel et l'autre moitié sans. La moitié des scénarios seront avec un bruit de fond de 75 dBA et l'autre moitié sans bruit. Un bruit de 75 dBA a environ la même intensité que le bruit d'une douche.

Chaque scénario se déroulera dans une salle avec différents bureaux. Vous serez attribué à un bureau de manière pseudo-aléatoire. Sur chaque bureau, il y a un ensemble de Lego. Vous devez suivre les instructions du mieux que vous pouvez, aussi rapidement que possible. Certaines instructions nécessiteront des informations ou des composants que les autres participants ont sur leurs bureaux. Vous devrez communiquer avec les autres participants pour les obtenir. Chaque fois que vous parlez à quelqu'un ou que vous écoutez quelqu'un, vous devez maintenir un contact visuel. Sinon, vous devriez vous concentrer sur votre question. Après un certain temps, on vous demandera de vous déplacer vers un autre bureau pour construire un autre ensemble Lego. Chaque scénario se terminera après 10 minutes. Les scénarios sont espacés par une pause de 5 minutes. Chaque scénario a un niveau sonore différent et un moyen de communication différent, plus de détails sont fournis ci-dessous.

A. Scenario sans effet d'environnement virtuel et sans bruit

Durant ce scénario, aucun bruit n'est diffusé directement dans votre oreille. Pour parler aux autres, vous devrez appuyer sur une touche du clavier qui vous sera fourni. Faites attention à appuyer sur la touche avant de commencer à parler, comme vous le feriez avec un talkie-walkie. Une fois que vous appuyez sur la touche, tout le monde vous entendra et ils ne pourront pas parler tant que vous maintiendrez la touche enfoncée. Vous devez relâcher la touche pour les entendre.

B. Scenario sans effet d'environnement virtuel et avec bruit

Durant ce scénario, un bruit sera diffusé directement dans vos oreilles. Le bruit est à 75 dBA. Pour parler aux autres, vous devrez appuyer sur une touche du clavier qui vous sera fourni. Faites attention à appuyer sur la touche avant de commencer à parler. Une fois que vous appuyez sur la touche, tout le monde vous entendra et ils ne pourront pas parler tant que vous maintiendrez la touche enfoncée. Vous devez relâcher la touche pour les entendre.

C. Scenario avec effet d'environnement virtuel et sans bruit

Durant ce scénario, aucun bruit n'est diffusé directement dans vos oreilles. Pour parler aux autres, vous n'avez pas besoin d'appuyer sur une touche. Vous devez ajuster votre voix et elle devrait être automatiquement détectée. L'intensité de votre voix est analysée pour créer un rayon autour de vous. Plus vous parlez fort, plus le rayon est grand, plus les gens peuvent vous entendre de loin. Vous devriez pouvoir parler naturellement, et les autres participants devraient pouvoir vous entendre. Vous pourriez avoir besoin d'ajuster votre voix pour bien atteindre les autres. De plus, la voix des autres participants aura une direction. Vous devriez avoir l'impression que leur voix vient de l'endroit où ils se trouvent.

D. Scenario avec effet d'environnement virtuel et avec bruit

Durant ce scénario, un bruit sera diffusé directement dans vos oreilles. Le bruit est à 75 dBA. Pour parler aux autres, vous n'avez pas besoin d'appuyer sur une touche. Vous devez toujours ajuster votre voix et elle devrait être automatiquement détectée.

Après chaque scénario, vous prendrez une courte pause de 5 minutes et on vous demandera de répondre à quelques questions brèves sur votre appréciation des technologies et des scénarios. Votre réponse n'affecte pas votre compensation.

3. Conclusion de l'expérience (5 mins)

Après l'expérience, le chercheur répondra à toute question supplémentaire et vous remettra la compensation. Vous signerez également un reçu pour la compensation.

III. Propriété des données

Nous vous rappelons que l'ensemble de vos données de recherche sera versé dans la banque de données RHAD Data, une banque de données ayant pour but de contribuer à faire avancer les connaissances dans le domaine de l'audiologie et de la protection auditive.

Par ailleurs, notez qu'en tout temps, vous demeurez propriétaire de vos données. Vos données ne seront utilisées qu'à des fins de recherche et ne seront jamais vendues.

Notez également que le chercheur responsable de la banque de données agira à titre de fiduciaire à l'égard des données qui seront conservées dans la banque. Il est donc responsable de la conservation, de la garde et de la sécurité des données, conformément au cadre de gestions de la banque. Le chercheur est également responsable de la distribution des données aux chercheurs qui en font la demande.

UTILISATION DES ENREGISTREMENTS

Le but premier des enregistrements audio et photos collectés dans le cadre du projet est de nous permettre de valider les différentes données.

Par ailleurs, avec votre consentement, ces enregistrements pourraient être utilisés à des fins d'enseignement, de recherche ou lors de conférences scientifiques.

Acceptez-vous que vos enregistrements soient utilisés à des fins d'étude, d'enseignement, de recherche ou lors de conférences scientifiques ? ☐ Oui ☐ Non

DÉCOUVERTE FORTUITE

Bien qu'ils ne fassent pas l'objet d'une évaluation médicale formelle, puisqu'il s'agit d'un projet de recherche, les résultats des tests, examens et procédures réalisés dans le cadre de ce projet de recherche peuvent mettre en évidence des problèmes jusque-là ignorés, c'est ce que l'on appelle une découverte fortuite. C'est pourquoi, en présence d'une particularité, la chercheure responsable du projet vous appellera pour assurer un suivi de l'information.

AVANTAGES ET BÉNÉFICES ASSOCIÉS AU PROJET DE RECHERCHE

Il se peut que vous retiriez un bénéfice personnel de votre participation à ce projet de recherche, mais nous ne pouvons vous l'assurer. Par ailleurs, les résultats obtenus contribueront à l'avancement des connaissances scientifiques dans ce domaine de recherche.

INCONVÉNIENTS ASSOCIÉS AU PROJET DE RECHERCHE

Cette expérience demande beaucoup de temps. Elle nécessite 1h30 de participation, dont environ 40 minutes nécessiteront votre participation. L'expérience se déroulera en deux séances. Vous pouvez ressentir de la fatigue. De plus, vous pouvez ressentir une gêne physique lorsque vous portez l'écouteur. Si l'inconfort physique est excessif et de longue durée, veuillez en informer l'expérimentateur et l'étude sera terminée. L'inconfort devrait s'atténuer après le retrait de l'écouteur.

RISQUES ASSOCIÉS AU PROJET DE RECHERCHE

Dans cette expérience, vous serez exposé à un niveau de 75 dBA (similaire au son d'une douche) pendant environ 20 minutes. Cette intensité est inférieure à la réglementation québécoise de 85 dBA. Vous ne pourrez pas ajuster le niveau d'exposition au bruit, mais il est sécuritaire car il est bien en dessous de la réglementation québécoise de 85 dBA. Si l'inconfort est trop intense, vous pourrez retirer les écouteurs et nous arrêterons le test. Vous ne pourrez pas ajuster le niveau d'exposition au bruit, car nous voulons avoir le même niveau d'exposition pour chaque participant.

Si, à tout moment, vous ressentez un inconfort, le test sera immédiatement interrompu, et vous pourrez choisir de continuer ou non l'expérience. Cependant, si vous avez été ou serez exposé à d'autres bruits significatifs le même jour (par exemple, travail dans un environnement très bruyant, participation à un concert, une manifestation, ...), votre exposition au bruit pendant la journée de votre participation au projet peut dépasser la dose quotidienne recommandée. Veuillez informer l'expérimentateur si vous avez de telles préoccupations, nous effectuerons alors les calculs pour évaluer votre niveau de risque.

Un autre risque possible est lié à votre identité. Étant donné que votre parole sera enregistrée, les personnes qui utilisent cette base de données pourraient vous reconnaître à travers votre voix.

PARTICIPATION VOLONTAIRE ET DROIT DE RETRAIT

Votre participation à ce projet de recherche est volontaire. Vous êtes donc libre de refuser d'y participer. Vous pouvez également vous retirer de ce projet à n'importe quel moment, sans avoir à donner de raisons, en informant l'équipe de recherche.

La chercheure responsable de ce projet de recherche, le Comité d'éthique de la recherche, l'École de technologie supérieure ou l'organisme subventionnaire peuvent mettre fin à votre participation, sans votre consentement. Cela peut se produire si de nouvelles découvertes ou informations indiquent que votre participation au projet n'est plus dans votre intérêt, si vous ne respectez pas les consignes du projet de recherche ou encore s'il existe des raisons administratives d'abandonner le projet.

Si vous vous retirez du projet ou êtes retiré(e) du projet, l'information et le matériel déjà recueillis dans le cadre de ce projet seront néanmoins conservés, analysés ou utilisés pour assurer l'intégrité du projet.

Toute nouvelle connaissance acquise durant le déroulement du projet qui pourrait avoir un impact sur votre décision de continuer à participer à ce projet vous sera communiquée rapidement.

CONTRIBUTION, CONSERVATION, ACCÈS À LA BANQUE DE DONNÉES ET CONFIDENTIALITÉ

Durant votre participation à ce projet de recherche, la chercheure responsable ainsi que les membres de l'équipe de recherche recueilleront, dans un dossier de recherche, les renseignements vous concernant et nécessaires pour répondre aux objectifs scientifiques du projet.

Ces renseignements peuvent comprendre votre nom, ainsi que les résultats de tous les tests et procédures réalisés dans le cadre du projet.

Tous les renseignements recueillis demeureront confidentiels, dans les limites prévues par la loi. Afin de préserver votre identité et la confidentialité de vos renseignements, un numéro de code vous sera attribué. La clé du code reliant votre nom à votre dossier de recherche sera conservée par la chercheure responsable de ce projet de recherche.

L'ensemble des renseignements vous concernant, le tout recueilli à titre de données de recherche, sera conservé de façon sécuritaire dans la banque de données constituée à des fins de recherche, logée à l'École de technologie supérieure, et ce, conformément au cadre de gestion de la banque.

Comme rappel, la banque de données établie à des fins de recherche par Jérémie Voix et Rachel Bouserhal permettra de développer une ressource mondiale pour favoriser la collaboration et développer de nouvelles connaissances dans le domaine de la santé et de la sécurité au travail.

Les données de recherche versées dans la banque de données seront partagées avec différents chercheurs. Ce transfert d'information implique que vos données de recherche pourraient être transmises dans d'autres pays que le Canada. Cependant, la chercheure responsable de ce projet de recherche respectera les règles de confidentialité en vigueur au Québec et au Canada, et ce, dans tous les pays.

Tous les projets de recherche utilisant les données de la banque seront évalués et approuvés par le Comité d'éthique de la recherche de l'École de technologie supérieure avant leur réalisation. Le Comité d'éthique de la recherche de l'École de technologie supérieure en assurera également le suivi.

Vos données de recherche seront conservées aussi longtemps qu'elles peuvent avoir une utilité pour l'avancement des connaissances scientifiques. Lorsqu'elles n'auront plus d'utilité, vos données de recherche seront détruites. Par ailleurs, notez qu'en tout temps, vous pouvez demander la non-utilisation de vos données de recherche en vous adressant à la chercheure responsable de ce projet de recherche. Dans une telle éventualité, vos données

de recherche déjà transférées à d'autres chercheurs seront néanmoins conservées, analysées ou utilisées, et ce, pour assurer l'intégrité des projets de recherche en cours et pour se conformer aux exigences règlementaires. Aucun nouveau projet ne sera réalisé avec vos données de recherche.

Les données de recherche pourront être publiées où faire l'objet de discussions scientifiques, mais il ne sera pas possible de vous identifier.

À des fins de surveillance, de contrôle, de protection, de sécurité, la banque de données ainsi que votre dossier de recherche pourront être consultés par une personne mandatée par des organismes réglementaires ainsi que par des représentants de l'organisme subventionnaire, de l'École de technologie supérieure ou du Comité d'éthique de la recherche. Ces personnes et ces organismes adhèrent à une politique de confidentialité.

Vous avez le droit de consulter votre dossier de recherche pour vérifier les renseignements recueillis et les faire rectifier au besoin.

COMPENSATION

En guise de compensation pour votre participation à ce projet de recherche, vous choisirez et recevrez l'une des 2 compensations ci-dessous : 20 dollars (en espèces) ou une paire de bouchons d'oreille de musicien (d'une valeur de 30 dollars). Si vous vous retirez du projet ou si votre participation est interrompue, vous ne recevrez pas la compensation.

POSSIBILITÉ DE COMMERCIALISATION

Votre participation au projet de recherche pourrait mener à la création de produits commerciaux qui pourraient être éventuellement protégés par voie de brevet ou autres droits de propriété intellectuelle. Cependant, dans un tel cas, vous ne pourrez en retirer aucun avantage financier.

EN CAS DE PRÉJUDICE

Si vous deviez subir quelque préjudice que ce soit dû à votre participation au projet de recherche, vous recevrez tous les soins et services requis par votre état de santé.

En acceptant de participer à ce projet de recherche, vous ne renoncez à aucun de vos droits, ni ne libérez la chercheure responsable, l'École de technologie supérieure et l'organisme subventionnaire de leur responsabilité civile et professionnelle.

PROCÉDURES EN CAS D'URGENCE MÉDICALE

L'École de technologie supérieure n'offre pas de services d'urgence. Par conséquent, advenant une condition médicale qui nécessiterait des soins immédiats, les premiers soins vous seront dispensés par le personnel en place et des dispositions seront prises afin de vous transférer, si nécessaire, aux urgences d'un hôpital avoisinant.

SUIVI ÉTHIQUE

Le comité d'éthique de la recherche de l'École de technologie supérieure a approuvé ce projet de recherche et en assure le suivi.

PERSONNES-RESSOURCES

Pour toute question en lien avec le projet de recherche, vous pouvez contacter Pascal Giard à l'adresse courriel « pascal.giard@etsmtl.ca ». Vous pouvez également contacter Rachel Bouserhal à l'adresse courriel « rachel.bouserhal@etsmtl.ca » ou Gabriel Ouaknine-Beaulieu à l'adresse courriel « gabriel.ouaknine-beaulieu.1@ens.etsmtl.ca ».

Pour toute question en lien avec vos droits en tant que participant à la recherche, vous pouvez contacter la coordonnatrice du Comité d'éthique de la recherche de l'École de technologie supérieure en téléphonant au (514) 396-8800 poste 7129.

CONSENTEMENT

Participant(e)

Je reconnais avoir lu le présent formulaire de consentement et avoir disposé de suffisamment de renseignements et du temps nécessaire pour prendre ma décision. Après réflexion, je consens volontairement à participer à ce projet de recherche, aux conditions énoncées.

Nom du (de la) participant(e)	*Signature*	*Date*

Personne qui obtient le consentement

J'ai expliqué au (à la) participant(e) tous les aspects pertinents de la recherche et j'ai répondu aux questions qu'il(elle) m'a posée.

Nom de la personne qui obtient le consentement	*Signature*	*Date*

Signature et engagement de la chercheure responsable de ce projet de recherche

Je certifie qu'on a expliqué au (à la) participant(e) le présent formulaire d'information et de consentement, que l'on a répondu aux questions qu'il(elle) avait.

Je m'engage, avec l'équipe de recherche, à respecter ce qui a été convenu au formulaire d'information et de consentement et à remettre une copie signée du présent formulaire au (à la) participant(e).

Chercheure responsable	*Signature*	*Date*

ÉCOLE DE TECHNOLOGIE SUPÉRIEURE
Université du Québec

CRITIAS
NSERC-EERS INDUSTRIAL RESEARCH CHAIR IN IN-EAR TECHNOLOGIES

H20231110

ICF – March 11, 2024 version

INFORMATION AND CONSENT FORM

TITLE OF THE RESEARCH PROJECT

Analysis of the added value of a communication system in a radio acoustic virtual environment (RAVE)

RESEARCHER IN CHARGE OF THE PROJECT

Pascal Giard, Professor in the Electrical Engineering Department– École de technologie supérieure (ÉTS)

CO-RESEARCHERS AND COLLABORATORS

Rachel Bouserhal, Professor in the Electrical Engineering Department – École de technologie supérieure (ÉTS)

Jérémie Voix, Professor in the Electrical Engineering Department – École de technologie supérieure (ÉTS)

Xinyi Zhang, PhD student in the Electrical Engineering Department – École de technologie supérieure (ÉTS)

Yves Méthot, Electronics Coordinator – CIRMMT

Julien Boissinot, Systems/Technical Manager – CIRMMT

Sylvain Pohu, Production Manager – CIRMMT

Valentin Pintat, Laboratory Manager at CRITIAS – École de technologie supérieure (ÉTS)

Dany Morissette, Research Assistant – École de technologie supérieure (ÉTS)

Arianne Marcoux, Research Assistant – École de technologie supérieure (ÉTS)

STUDENT

Gabriel Ouaknine-Beaulieu, Master's student in the Department of Electrical Engineering – École de technologie supérieure (ÉTS)

FUNDING

Pascal Giard received funding from Natural Science and Engineering Research Council of Canada (NSERC) as part of the CRITIAS Chair to carry out the research project.

INTRODUCTION

We are inviting you to take part in a research project. However, before accepting to participate in the project and signing this Information and Consent Form, please take the time to read, understand, and carefully consider the following information.

This form contains words that you may not understand. We invite you to ask any questions that you feel may be useful to the researcher in charge of the project or a member of the research team and to ask them to explain any words or expressions that are unclear to you.

NATURE AND OBJECTIVES OF THE RESEARCH PROJECT

The objective of this research project is to evaluate the added value of a radio acoustic virtual environment in a communication system.

The research project also aims to contribute to a data bank established for the purpose of research and entitled *CRITIAS : DB*. The goal of this data bank is to group together all data collected by various projects carried out by Professor Jérémie Voix and Rachel Bousheral and to make them available to researchers at the École de technologie supérieure and other institutions in an effort to advance knowledge in the field of occupational health and safety.

To carry out the project, we intend to recruit 30 participants, all genders, aged 18 years or older.

EXECUTION OF THE RESEARCH PROJECT

I. Location and Duration of the Participation

Your total participation in the project will last 1 hour and 30 minutes and require 2 visits.

First visit (30 minutes) : a preselection will take place at the Chaire de recherche industrielle ÉTS-EERS en technologies intra-auriculaires (CRITIAS) at 355 Peel, Montréal.

Second visit (1 hour) : The data collection will take place at McGill University in the Centre for Interdisciplinary Research in Music Media and Technology (CIRMMT) at 527 Sherbrooke Ouest, Montréal.

II. Nature of your Participation

1. Eligibility assessment (about 30 min)

First, your eligibility for participation will be assessed at CRITIAS. You will fill out an eligibility form and complete an otoscopic inspection. An otoscopic inspection assesses the status of your outer ear and your eardrum. Afterward, you'll perform a pure-tone audiometry test. This test intends to assess the sensitivity of your hearing. During this test, you'll hear different sounds at different intensities and different frequencies. You'll have to answer on a tablet if you can hear the sound or not. At the end of this step, it is possible that you may not be able to proceed with further steps. If this is the case, the researcher will explain the reasons. If you are eligible, the width of your head, the space between your ears and your face, as well as some other measurements of your face, will be taken to select audio filters that will be as personalized as possible for you, using a machine learning algorithm.

	
Fig. 1 Otoscope & tips	Fig. 2 Pure-tone audiometry test kit

2. Data collection (55 mins)

The data collection will be at CIRMMT. You will be paired with two other participants and together you will take place in four scenarios of ten minutes. In each scenario, you will wear an earpiece, and a hard hat with a tracking device. Those devices will allow you to communicate with the other participants and create the virtual environment effect. You will also wear a belt that records your heartbeat. This belt will be used as an impartial measurement of your appreciation. Half the scenarios will be with the virtual environment effect and the other half will be without. Half the scenarios will be with a background noise of 75 dBA and half without noise. A 75 dBA noise is about the same intensity as the noise of a shower.

Each scenario will be in a room with different desks. You'll be assigned a desk pseudo-randomly. On each desk, there's a Lego set. You must follow the instructions to the best of your ability, as fast as you can. Some instructions will need information or components that the other participants have on their desks. You will need to communicate with the other participants to find out. Every time you're speaking with someone or listening to someone, you need to keep visual contact. If not, you should be focused on your question. After some time, you'll be asked to move to another desk to build another Lego set. Each scenario will end after 10 minutes. The scenarios are spaced by a 5-minute break. Each scenario has a different noise level and a different means to communicate, more details are provided in the following.

A. Scenario without noise and without virtual environment effect

During this scenario, no noise is played directly in your ear. To speak with the others, you will need to press a key on the keyboard given to you. You should be careful of pressing the key before starting to speak, as you would do using a walkie talkie. Once you press the key, everyone will hear you and they won't be able to speak as long as you press the key. You need to release the key to hear them.

B. Scenario With noise and without virtual environment effect

During this scenario, a noise will be played directly into your ears. The noise is at 75 dBA. To speak with the others, you will need to press a key on the keyboard given to you. You should be careful of pressing the key before starting to speak. Once you press the key, everyone will hear you and they won't be able to speak as long as you press the key. You need to release the key to hear them.

C. Scenario without noise and With virtual environment effect

During this scenario, no noise is played directly in your ears. To speak with the others, you don't need to press any key. You need to adjust your voice and it should automatically detect your voice. The intensity of your voice is analyzed to create a radius surrounding you. The louder you speak, the bigger the radius, the further people can

hear you. You should be able to speak naturally, and the other participants should be able to hear you. You may need to adjust your voice to reach properly the others. In addition, the voice coming from the other participants will have a direction. You should have the impression their voice is coming from where they are.

D. Scenario With noise and With virtual environment effect

During this scenario, a noise will be played directly into your ears. The noise is at 75 dBA. To speak with the others, you don't need to press any key. You still need to adjust your voice and it should automatically detect your voice.

After each scenario, you will take a short 5-minute break and you'll be asked to answer some short questions on your appreciation of the technologies and the scenarios. Your answer does not affect your compensation.

3. Experimentation conclusion (5 min)

Following the experimentation, the experimenter will answer any additional questions and provide you with compensation. You will also sign a receipt of the compensation.

III. Ownership of Data

You will remain the sole owner of your data at all times. It will only be used for the purpose of research and will never be sold.

The researchers in charge of the data bank will act as the trustees of all data in the data bank. He/she will be responsible for maintaining the data, for its stewardship and its security in accordance with the Data Bank Governance Framework. The researchers will also be responsible for sharing data with researchers who request it.

USE OF RECORDINGS

The primary goal of the audio and other recordings collected as part of the project is to allow us to validate the data.

As such, with your consent, these recordings may be used for education, research, or scientific lectures.

Do you consent to your recordings being used for education, research, or scientific lectures?
☐ Yes ☐ No

INCIDENTAL FINDINGS

Although they are not subject to a formal medical assessment since they will be part of a research project, the results of all tests, examinations, and procedures performed as part of this research project may reveal problems that were unknown until now, which are referred to as incidental findings. As a result, if a peculiar observation is brought to light, the researcher in charge of the project will inform you to ensure follow-up.

ADVANTAGES ASSOCIATED WITH THE RESEARCH PROJECT

You may gain personal benefit from participating in this research project. However, we cannot ensure that this will be the case. The results obtained will nonetheless contribute to advancing scientific knowledge in this field of research.

Inconveniences Associated With the Research Project

This research requires 1 hour 30 min of participation, of which about 40 minutes will require your active participation. The experience will take place in two sessions. You may experience fatigue. In addition, you may experience physical discomfort when wearing the earpiece. If the physical discomfort is excessive and long-lasting, please let the experimenter know and the study will be terminated. The discomfort is expected to subside after the removal of the earpiece.

Risks Associated With the Research Project

In this experiment, you will be exposed to a level of 75 dBA (similar to the sound of a shower) for around 20 mins. This intensity is below Quebec's regulation of 85dBA. You won't be able to adjust the noise exposure level, but it is safe as it is well below Quebec's regulation of 85dBA. If the discomfort is too intense, you will be able to take out the earpieces and we will stop the test. You won't be able to adjust the noise exposure level, as we want to have the same noise exposure level for every participant.

If, at any time, you feel discomfort, the test will be stopped immediately, and you will be able to choose whether or not to continue the experiment.

However, if you have been or will be exposed to other significant noise on the same day (e.g., work in a very noisy environment, attending a concert, a protest, ...), your exposure to noise during the day of your participation in the project may exceed the recommended daily noise dose. Please let the experimenter know if you have such concerns, we will then perform the calculations to assess your risk level.

Another possible risk is related to your identity. Since your speech will be recorded, people who use this database may recognize you through your voice.

Voluntary Participation and Right to Withdraw

Your participation in this research project is voluntary. You are therefore free to refuse to participate. You may also withdraw from the project at any time without the need to provide any reason, by informing the research team.

The researcher in charge of the project, the Research Ethics Committee, the École de technologie supérieure, or the funding agency may decide to end your participation without your consent. This can happen if discoveries or information reveal that your participation in the project is no longer in your interest, or if you do not follow the research project instructions, or if there are administrative reasons for abandoning the project.

If you withdraw or are withdrawn from the project, the information and material already collected as part of the project will still be held on file, analyzed, or used to ensure the integrity of the project.

All new knowledge acquired during the project that may impact your decision to continue to participate will be shared with you in short order.

Contribution, Retention, Access to the Data Bank, and Confidentiality

During your participation in this project, the researcher in charge of the project and the members of the research team will collect and record the information about you in a research file. They will only collect the information required to achieve the scientific objectives of the project.

This information may include your name, your age, your history of ear surgery along with the results of all tests and procedures carried out as part of this project.

All of the information collected will remain confidential, within the limits set out under the law. In order to protect your identity and the confidentiality of your information, you will be assigned a code number. The researcher in charge of the research project will keep the key code linking your name to the research file.

All of your information, all collected as research data, will be retained in a secure manner in the data bank established for the purpose of research at the École de technologie supérieure, in accordance with the Data Bank Governance Framework.

As a reminder, the data bank established for the purpose of research by Jérémie Voix and Rachel Bouserhal will allow to develop a global resource to foster collaboration and develop new knowledge in the field of occupational health and safety.

The research data retained in the data bank will be shared with other researchers. This means that your research data may be transferred in countries other than Canada. However, the researcher in charge of this data bank will follow the confidentiality regulations in effect in Quebec and in Canada, regardless of the country.

All research projects that use the data bank will be evaluated and approved by the Research Ethics Committee of the École de technologie supérieure prior to being carried out. The Committee will also follow up with these projects.

Your research data will be retained as long as it remains useful to the advancement of scientific knowledge. When it is no longer useful, your research data will be destroyed. Please note that at any time, you may request that your research data not be used by contacting the researcher in charge of the project. In such a case, your research data that has already been transferred to other researchers will still be maintained, analyzed, or used to protect the integrity of research projects that are already underway and to comply with regulatory requirements. However, your data will not be used for any new projects.

The research data may also be published or used for scientific discussions. However, it will not be possible to identify you.

For the purpose of monitoring, control, protection, and security, your research file may be consulted by individuals authorized by regulatory organizations, representatives of the funding agency, the École de technologie supérieure, or the Research Ethics Committee. These individuals and organizations are bound by a confidentiality policy.

You have the right to consult your research file to verify the information collected and to modify it as needed.

COMPENSATION

As compensation for your participation, you will choose and receive one of the 2 compensations below 20 dollars (cash) or a pair of musician earplugs (worth 30 dollars). If you withdraw from the project or your participation is interrupted, you will not receive the compensation.

POSSIBILITY OF MARKETING

Your participation in this research project may lead to the creation of commercial products that may eventually be protected under a patent or other intellectual property rights. In such a case, you will not receive any resulting financial benefit.

In Case of Prejudice

Should you suffer any prejudice due to your participation in the research project, you will receive all the care and services required by your state of health.

By accepting to participate in this research project, you are not waiving any of your legal rights or releasing the researcher in charge of the project, the École de technologie supérieure, and the funding agency from their civil and professional responsibilities.

Procedures in the Event of a Medical Emergency

Please note that the École de technologie supérieure does not offer emergency services. Therefore, in the event of a medical condition that requires immediate care, first aid will be provided to you by the personnel on site and arrangements will be made to transfer you, if necessary, to the emergency room of a nearby hospital.

Monitoring of Ethical Aspects

The Research Ethics Committee of the École de technologie supérieure approved this project and will ensure the follow-up.

Contact Persons

If you have any questions regarding the research project, you can contact Pascal Giard at pascal.giard@etsmtl.ca. You may also contact Rachel Bouserhal at rachel.bouserhal@etsmtl.ca or Gabriel Ouaknine-Beaulieu at gabriel.ouaknine-beaulieu.1@ens.etsmtl.ca.

For all questions regarding your rights as a participant in the research, you may contact the Research Ethics Committee Coordinator of the École de technologie supérieure at 514-396-8800, ext. 7129.

CONSENT

Participant

I acknowledge that I have read this consent form and have been provided with all the information and enough time to make my decision. Upon reflection, I voluntarily consent to participate in this research project in keeping with the conditions set out herein.

Name of the participant	*Signature*	*Date*

Person Obtaining Consent (if other than the researcher in charge of the research project)

I explained all relevant aspects of the research to the participant and answered all of the questions that he/she asked.

Name of the Individual Obtaining Consent	*Signature*	*Date*

Signature and Commitment of the Researcher in Charge of the Project

I hereby certify that we have explained this Information and Consent Form to the participant and answered all of their questions.

I agree, along with the research team, to respect everything that was agreed to in the information and consent form and to give a signed copy of this form to the participant.

Researcher in charge of the project	*Signature*	*Date*

BIBLIOGRAPHIE

Abel, S. M., Alberti, P. W., Haythornthwaite, C. & Riko, K. (1982). Speech intelligibility in noise : Effects of fluency and hearing protector type. *J. Acoust. Soc. Am.*, 71(3), 708-715. doi : 10.1121/1.387547.

Abel, S. M., Armstrong, N. M. & Giguère, C. (1991). Signal detection and speech perception with level-dependent hearing protectors. *Canadian Acoustics*, 19(4), 81-82.

Abel, S., Tsang, S. & Boyne, S. (2007). Sound localization with communications headsets : Comparison of passive and active systems. *Noise Health*, 9(37), 101. doi : 10.4103/1463-1741.37426.

Acoustical Society of America. (2001). *Design Response of Weighting Networks for Acoustical Measurements*. ANSI S1.42-2001. Melville, New York, États-unis : American National Standards Institute.

Acoustical Society of America. (2009). *Methods for Manual Pure-Tone Threshold Audiometry*. ANSI S3.21-2004. Melville, New York, États-unis : American National Standards Institute.

Ashmore, J., Avan, P., Brownell, W., Dallos, P., Dierkes, K., Fettiplace, R., Grosh, K., Hackney, C., Hudspeth, A., Jülicher, F., Lindner, B., Martin, P., Meaud, J., Petit, C., Santos Sacchi, J. & Canlon, B. (2010). The remarkable cochlear amplifier. *Hearing Research*, 266(1), 1-17. doi : https://doi.org/10.1016/j.heares.2010.05.001.

Bellows, S. & Leishman, T. (2022, September). Effect of Head Orientation on Speech Directivity. *Proc. Conf. Int. Speech Commun. Assoc. (Interspeech)*, 23, 246–250. doi : 10.21437/Interspeech.2022-553.

Bitzer, J., Bilert, S. & Holube, I. (2018). Evaluation of binaural own voice detection (OVD) algorithms. *Speech Communication ; 13th ITG-Symposium*, pp. 1–5.

Bonnet, F., Nelisse, H., Nogarolli, M. & Voix, J. (2019). In-ear noise dosimetry under earplug : Method to exclude wearer-induced disturbances. *Int. J. Industr. Ergonom.*, 74, 139–148. doi : 10.1016/j.ergon.2019.102862.

Bou Serhal, R. E., Falk, T. H. & Voix, J. (2013). Integration of a distance sensitive wireless communication protocol to hearing protectors equipped with in-ear microphones. *Proc. Meet. Acoust.*, 19(1), 040013.

Bouserhal, R. E., Bockstael, A., MacDonald, E., Falk, T. H. & Voix, J. (2017a). Modeling Speech Level as a Function of Background Noise Level and Talker-to-Listener Distance for Talkers Wearing Hearing Protection Devices. *J. Speech Lang. Hear. Res.*, 60(12), 3393–3403. doi : 10.1044/2017_JSLHR-S-17-0052.

Bouserhal, R. E., Falk, T. H. & Voix, J. (2017b). In-ear microphone speech quality enhancement via adaptive filtering and artificial bandwidth extension. *J. Acoust. Soc. Am.*, 141(3), 1321–1331. doi : 10.1121/1.4976051.

Brammer, A. J., Yu, G., Bernstein, E. R., Cherniack, M. G., Peterson, D. R. & Tufts, J. B. (2014a). Understanding speech when wearing communication headsets and hearing protectors with subband processinga). *J. Acoust. Soc. Am.*, 136(2), 671-681. doi : 10.1121/1.4883385.

Brammer, A. J., Yu, G., Bernstein, E. R., Cherniack, M. G., Peterson, D. R. & Tufts, J. B. (2014b). Understanding speech when wearing communication headsets and hearing protectors with subband processinga). *J. Acoust. Soc. Am.*, 136(2), 671-681. doi : 10.1121/1.4883385.

Brinkmann, F., Dinakaran, M., Pelzer, R., Wohlgemuth, J. J., Seipel, F., Voss, D., Grosche, P. & Weinzierl, S. [Accessed : 2025-06-18]. (2019). The HUTUBS Head-Related Transfer Function (HRTF) Database. Technische Universität Berlin.

Carillo, K., Doutres, O. & Sgard, F. (2020). Theoretical investigation of the low frequency fundamental mechanism of the objective occlusion effect induced by bone-conducted stimulation. 147(5), 3476-3489. doi : 10.1121/10.0001237.

Casali, J. (2010). Powered Electronic Augmentations in Hearing Protection Technology Circa 2010 including Active Noise Reduction, Electronically-Modulated Sound Transmission, and Tactical Communications Devices : Review of Design, Testing, and Research. *Int. J. Acoust. Vib.*, 15, 168-186. doi : 10.20855/ijav.2010.15.4269.

Casali, J. G. & Gerges, S. N. Y. (2006). Protection and Enhancement of Hearing in Noise. *Reviews of Human Factors and Ergonomics*, 2(1), 195-240. doi : 10.1177/1557234X0600200108.

Casali, J. G. & Horylev, M. J. (1987). Speech discrimination in noise : The influence of hearing protection. *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, 31(11), 1246–1250. doi : 10.1177/154193128703101116.

Casali, J. G. & Tufts, J. B. (2022). Auditory Situation Awareness and Speech Communication in Noise. Dans Deanna.K.Meinke, Elliott.H.Berger, Neitzel, R., Driscoll, D. & Kathryn.Bright (Éds.), *The Noise Manual* (éd. 6, pp. 419-458). 3141 Fairview Park Drive, Suite 777 Falls Church, VA 22042 : the American Industrial Hygiene Association.

Casto, K. & Casali, J. (2013). Effects of Headset, Flight Workload, Hearing Ability, and Communications Message Quality on Pilot Performance. *Hum. Factors*, 55, 486-98. doi : 10.1177/0018720812461013.

C.Bielefeld, E. (2022). Anatomy and Physiology of the Ear : Normal Function and the Damage Underlying Hearing Loss. Dans Deanna.K.Meinke, Elliott.H.Berger, Neitzel, R., Driscoll, D. & Kathryn.Bright (Éds.), *The Noise Manual* (éd. 6, pp. 67-78). 3141 Fairview Park Drive, Suite 777 Falls Church, VA 22042 : the American Industrial Hygiene Association.

Chen, J., Benesty, J., Huang, Y. & Diethorn, E. (2008). Fundamentals of Noise Reduction. Dans Benesty, J., Sondhi, M. & Huang, Y. (Éds.), *Springer Handbook of Speech Processing* (éd. 1, pp. 843-871). Berlin Heidelberg, Germany : Springer.

Christensen, R. H. B. [R package version 2023.12-4.1]. (2023). ordinal—Regression Models for Ordinal Data [Manual]. Repéré à https://CRAN.R-project.org/package=ordinal.

Driscoll, D. P. (2022). Physics of Sound and Vibration. Dans Deanna.K.Meinke, Elliott.H.Berger, Neitzel, R., Driscoll, D. & Kathryn.Bright (Éds.), *The Noise Manual* (éd. 6, pp. 11-27). 3141 Fairview Park Drive, Suite 777 Falls Church, VA 22042 : the American Industrial Hygiene Association.

Dunn, H. K. & Farnsworth, D. W. (1939). Exploration of Pressure Field Around the Human Head During Speech. *J. Acoust. Soc. Am.*, 10(3), 184-199. doi : 10.1121/1.1915975.

Elliott, T. M. & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS computational biology*, 5(3), e1000302. doi : 10.1371/journal.pcbi.1000302.

Fostick, L. & Fink, N. (2021). Situational awareness : The effect of stimulus type and hearing protection on sound localization. *Sensors*, 21(21), 7044.

Fujimoto, M. & Ishizuka, K. (2008). Noise Robust Voice Activity Detection Based on Switching Kalman Filter. *IEICE Trans.*, 91-D, 467-477. doi : 10.1093/ietisy/e91-d.3.467.

Fux, T. (2012). *Vers un système indiquant la distance d'un locuteur par transformation de sa voix*. (Thèse de doctorat, Université de Grenoble, 5 rue du Général Cassagnou, 68300, Saint-Louis, France).

Gallagher, H. L., McKinley, R. L., Theis, M. A., Swayne, B. J. & Thompson, E. R. (2014). *Performance Assessment of Active Hearing Protection Devices :*.

Gaston, J., Foots, A., Mermagen, T. & Scharine, A. (2019). The effectiveness of tactical communication and protection systems (TCAPS) on minimizing hearing hazard and maintaining auditory situational awareness. *Advances in Human Factors in Simulation and Modeling : Proceedings of the AHFE 2018 International Conferences on Human Factors and Simulation and Digital Human Modeling and Applied Optimization*, 780, 382-391.

Giguere, C., Behar, A., Dajani, H. R., Kelsall, T. & Keith, S. E. (2012). Direct and indirect methods for the measurement of occupational sound exposure from communication headsets. *Noise Control Engineering Journal*, 60(6), 630-644.

Giguère, C. & Dajani, H. R. (2009). Noise exposure from communications headsets : The effects of external noise, device attenuation and effective listening signal-to-noise ratio. *INTER-NOISE and NOISE-CON Congress and Conference Proceedings*, 2009(6), 1005-1012.

Giguère, C., Laroche, C., Vaillancourt, V. & Soli, S. (2010). Modelling Speech Intelligibility in the Noisy Workplace for Normal-hearing and Hearing-impaired Listeners Using Hearing Protectors. *Int. J. Acoust. Vib.*, 15, 156-167. doi : 10.20855/ijav.2010.15.4268.

Giguère, C., Vaziri, G., Dajani, H. R. & Berger, E. H. (2017). Speech communication with hearing protectors. *National Hearing Conservation Association Spectrum*, 34, 15-19.

H. Berger, E. & Voix, J. (2021). Hearing Protection Devices. Dans *The Noise Manual* (éd. 6th edition, pp. 380). the American Industrial Hygiene Association.

Houtgast, T., Steeneken, H. J. & Plomp, R. (1980). Predicting speech intelligibility in rooms from the modulation transfer function. I. General room acoustics. *Acta Acustica united with Acustica*, 46(1), 60-72.

Howell, K. & Martin, A. (1975). An investigation of the effects of hearing protectors on vocal communication in noise. *J. Sound. Vib.*, 41(2), 181-196. doi : https://doi.org/10.1016/S0022-460X(75)80096-4.

Jokinen, E., Yrttiaho, S., Pulakka, H., Vainio, M. & Alku, P. (2012). Signal-to-noise ratio adaptive post-filtering method for intelligibility enhancement of telephone speech. *J. Acoust. Soc. Am.*, 132(6), 3990-4001. doi : 10.1121/1.4765074.

Kim, G. & Cho, N. I. (2006). Two-microphone voice activity detection in the presence of coherent interference. *Proc. Int. Conf. Spoken Lang. Process. (INTERSPEECH–ICSLP)*, 9, 1686–1689. doi : 10.21437/Interspeech.2006-469.

Kryter, K. D. (1946). Effects of Ear Protective Devices on the Intelligibility of Speech in Noise. *J. Acoust. Soc. Am.*, 18(2), 413-417. doi : 10.1121/1.1916380.

Laroche, C., Giguère, C., Vaillancourt, V., Marleau, C., Cadieux, M.-F., Laprise-Girard, K., Gula, E., Carroll, V., Bibeau, M. & Nélisse, H. (2022). Effect of Hearing and Head Protection on the Localization of Tonal and Broadband Reverse Alarms. *Hum. Factors*, 64(7), 1105–1120. doi : 10.1177/0018720821992223.

Lee, K. & Casali, J. G. (2017). Development of an auditory situation awareness test battery for advanced hearing protectors and TCAPS : detection subtest of DRILCOM (detection-recognition/identification-localization-communication). *International journal of audiology*, 56(sup1), 22–33.

LETOWSKI, T. & MCGEE, L. (1993). Detection of warble tones in wideband noise with and without hearing protection devices. *The Annals of occupational hygiene*, 37(6), 607–614.

Lezzoum, N., Gagnon, G. & Voix, J. (2014). Voice Activity Detection System for Smart Earphones. *IEEE Trans. Consum. Electron.*, 60, 737-744. doi : 10.1109/TCE.2014.7027350.

Lezzoum, N., Gagnon, G. & Voix, J. (2016a). Noise reduction of speech signals using time-varying and multi-band adaptive gain control for smart digital hearing protectors. *Appl. Acoust.*, 109, 37-43. doi : https://doi.org/10.1016/j.apacoust.2016.03.001.

Lezzoum, N., Gagnon, G. & Voix, J. (2016b). Echo threshold between passive and electroacoustic transmission paths in digital hearing protection devices. *Int. J. Industr. Ergonom.*, 53, 372-379. doi : https://doi.org/10.1016/j.ergon.2016.04.004.

Lindau, A., Hohn, T. & Weinzierl, S. (2007). Binaural Resynthesis for Comparative Studies of Acoustical Environments. *Proc. Audio Eng. Soc.*, 122, 69-87.

Lindeman, H. E. (1976). Speech Intelligibility and the Use of Hearing Protectors. *Int. J. Audiol.*, 15(4), 348–356. doi : 10.3109/00206097609071794.

Majdak, P., Iwaya, Y., Carpentier, T., Nicol, R., Parmentier, M., Roginska, A., Suzuki, Y., Watanabe, K., Wierstorf, H., Ziegelwanger, H. & Noisternig, M. (2013, 05). Spatially Oriented Format for Acoustics : A Data Exchange Format Representing Head-Related Transfer Functions. *Audio Eng. Soc. Conv.*, 134, 1-11.

McBride, M., Weatherless, R., Mermagen, T. & Letowski, T. (2009). Effects of hearing protection on speech communication. *16th International Congress on Sound and Vibration 2009, ICSV 2009*, 1(16), 178-185.

Ministère du Travail, de l'Emploi et de la Solidarité sociale. (2013). *Règlement sur la santé et la sécurité du travail*. Loi sur la santé et la sécurité du travail (chapitre S-2.1, a. 223). Publications Québec.

Murphy, W. J., Kardous, C. A. & Brueck, S. E. (2022). Sound Measurement : Instrumentation and Noise Metrics. Dans Deanna.K.Meinke, Elliott.H.Berger, Neitzel, R., Driscoll, D. & Kathryn.Bright (Éds.), *The Noise Manual* (éd. 6, pp. 29-66). 3141 Fairview Park Drive, Suite 777 Falls Church, VA 22042 : the American Industrial Hygiene Association.

Niermann, C. (2015, 09). Can spatial audio support pilots ? 3D-audio for future pilot-assistance systems. *IEEE/AIAA Digital Avionics Syst. Conf. (DASC)*, pp. 3C5-1. doi : 10.1109/DASC.2015.7311401.

O'Shaughnessy, D. (2008). Formant Estimation and Tracking. Dans Benesty, J., Sondhi, M. & Huang, Y. (Éds.), *Springer Handbook of Speech Processing* (éd. 1, pp. 213-227). Berlin Heidelberg, Germany : Springer.

Paliwal, K. K., Lyons, J. G. & Wójcicki, K. K. (2010). Preference for 20-40 ms window duration in speech analysis. *Int. Conf. on Signal Process. and Commun. Syst.*, pp. 1-4. doi : 10.1109/ICSPCS.2010.5709770.

Pang, J. (2017, Jan). Spectrum Energy Based Voice Activity Detection. *Proc. IEEE. Comput. Commun. Workshop Conf. (CCWC)*, 7, 9–13. doi : 10.1109/CCWC.2017.7868454.

Park, H., Shin, Y.-S. & Shin, S.-H. (2019). Speech Quality Enhancement for In-Ear Microphone Based on Neural Network. *IEICE Trans. Inf. Syst.*, E102.D, 1594-1597. doi : 10.1587/transinf.2018EDL8249.

Pawełczyk, M., Latos, M., Michalczyk, M., Czyz, K. & Mazur, K. (2011). An efficient communication system for noisy environments. *Proc. Int. Conf. Model. Identif. Control*, pp. 273-277. doi : 10.1109/ICMIC.2011.5973714.

Pelegrín-García, D., Smits, B., Brunskog, J. & Jeong, C.-H. (2011). Vocal effort with changing talker-to-listener distance in different acoustic environments. *J. Acoust. Soc. Am.*, 129(4), 1981–1990. doi : 10.1121/1.3552881.

Pelzer, R., Dinakaran, M., Brinkmann, F., Lepa, S., Grosche, P. & Weinzierl, S. (2020). Head-related transfer function recommendation based on perceptual similarities and anthropometric features. *J. Acoust. Soc. Am.*, 148(6), 3809-3817. doi : 10.1121/10.0002884.

Pertilä, P., Fagerlund, E., Huttunen, A. & Myllyla, V. (2021). Online Own Voice Detection for a Multi-Channel Multi-Sensor In-Ear Device. *IEEE Sensors J.*, PP, 1-1. doi : 10.1109/JSEN.2021.3122936.

R Core Team. (2022). R : A Language and Environment for Statistical Computing. Repéré à https://www.R-project.org/.

Ribera, J. E., Mozo, B. T. & Murphy, B. A. (2004). Speech intelligibility with helicopter noise : tests of three helmet-mounted communication systems. *Aviat. Space Environ. Med.*, 75(2), 132–137. doi : 10.3357/ASEM.1113.2004.

Same, M., Gandubert, G., Gleeton, G., Ivanov, P. & Landry, R. (2020). Simplified Welch Algorithm for Spectrum Monitoring. *Appl. Sci.*, 11, 86. doi : 10.3390/app11010086.

Sehgal, A. & Kehtarnavaz, N. (2018). A Convolutional Neural Network Smartphone App for Real-Time Voice Activity Detection. *IEEE Access*, PP, 1-1. doi : 10.1109/ACCESS.2018.2800728.

Seris, J., Gargour, C. & Laville, F. (2007). VAD INNES : A Voice Activity Detector for Noisy Industrial Environments. *Proc. IEEE Int. Midwest Symp. Circuits Syst. (MWSCAS)*, pp. 377–380. doi : 10.1109/MWSCAS.2007.4488609.

Smyth, T. (2019). Music 175 : Time and Space [Notes de cours]. Repéré à http://musicweb.ucsd.edu/~trsmyth/space175/space175.pdf.

Sohn, J., Kim, N. S. & Sung, W. (1999). A statistical model-based voice activity detection. *IEEE Signal Process. Lett.*, 6(1), 1-3. doi : 10.1109/97.736233.

Sohn, J. & Sung, W. (1998). A voice activity detector employing soft decision based noise spectrum adaptation. *IEEE Int. Conf. on Acoustics, Speech, and Signal Process. (ICASSP)*, 1, 365–368. doi : 10.1109/ICASSP.1998.674443.

Traunmüller, H. & Eriksson, A. (2000). Acoustic effects of variation in vocal effort by men, women, and children. *J. Acoust. Soc. Am.*, 107(6), 3438-3451. doi : 10.1121/1.429414.

Urquhart, R. L. & Casali, J. G. (2005). Communications in Severe Low-Frequency Noise : An Investigation of Microphone Type and Speech Level on Intelligibility and Attenuation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 49(19), 1771-1775.

Valentin, O., Ezzaf, S., Gauthier, P.-A., Berbiche, D., Negrini, A., Doutres, O., Sgard, F. & Berry, A. (2024). Assessing the multidimensional comfort of earplugs in virtual industrial noise environments. *Applied Ergonomics*, 121, 104343.

Varga, A. & J.M.Steeneken, H. (1993). Assessment for automatic speech recognition : II. NOISEX-92 : A database and an experiment to study the effect of additive noise on speech recognition systems. *Speech Commun.*, 12(3), 247-251. doi : https://doi.org/10.1016/0167-6393(93)90095-3.

Voix, J. & Laville, F. (2009). The objective measurement of individual earplug field performance. *J. Acoust. Soc. Am.*, 125(6), 3722–3732.

Westerlund, N., Dahl, M. & Claesson, I. (2005). *In-Ear Microphone Techniques for Severe Noise Situations* (Rapport n°2005 :12).

Zhu, Z., Zhang, L., Pei, K. & Chen, S. (2023). A Robust and Lightweight Voice Activity Detection Algorithm for speech enhancement at Low Signal-to-noise Ratio. *Digit. Signal Process.*, 141, 104151. doi : 10.1016/j.dsp.2023.104151.