Toucher social en robotique de téléprésence ubiquïte : imbrication des facteurs physiques et socio-affectifs dans la portée vocale en interaction

Ambre Davat

Université Grenoble Alpes

THÈSE

Pour obtenir le grade de
DOCTEUR DE L'UNIVERSITÉ GRENOBLE ALPES
Spécialité : **SIGNAL, IMAGE, PAROLE, TELECOMS**

Arrêté ministériel : 25 mai 2016

Présentée par
Ambre DAVAT

Thèse dirigée par **Gang FENG,**
Professeur de l'Université Grenoble Alpes
et codirigée par **Véronique AUBERGÉ,**
Chargée de recherche CNRS

préparée au sein des laboratoires **GIPSA-lab** et **LIG**
dans **l'École Doctorale Électronique, Électrotechnique, Automatique et Traitement du Signal (EEATS)**

Toucher social en robotique de téléprésence ubiquïte :
imbrication des facteurs physiques et socio-affectifs dans la portée vocale en interaction

Thèse soutenue publiquement le **13 novembre 2020**,
devant le jury composé de :

M. Laurent BESACIER
Professeur, Université de Grenoble – LIG, Président
M. Jean-François BONASTRE
Professeur, Université d'Avignon – LIA, Rapporteur
M. Mohamed CHETOUANI
Professeur, Sorbonne Université – ISIR, Rapporteur
Mme. Fabienne MARTIN-JUCHAT
Professeure, Université de Grenoble – GRESEC, Examinatrice
M. Gang FENG
Professeur, Université de Grenoble – GIPSA-lab, Directeur de thèse
Mme. Véronique AUBERGÉ
Chargée de recherche CNRS, Université de Grenoble – LIG, Co-directrice de thèse

« Et c'est ainsi qu'on érode les montagnes.
De l'eau qui goutte sur une pierre, qui la dissout et l'élimine.
Ainsi qu'on change la face du monde, une goutte à la fois.
De l'eau qui goutte sur une pierre, commissaire.
De l'eau qui circule sous terre,
qui remonte en bouillonnant
là où on ne l'attend pas. »

TERRY PRATCHETT
Les Annales du disque-monde
Jeu de nains
traduit de l'anglais par PATRICK COUTON

(Librairie de l'Atalante, 2008)

REMERCIEMENTS

Je tiens à faire part de ma gratitude à toutes les personnes qui m'ont accompagnée pendant ces quatre années de thèse.

Tout d'abord, je remercie mes encadrants. Concilier vos deux approches n'était pas toujours facile, mais m'a apporté bien plus humainement et intellectuellement que ce que je peux exprimer dans ces quelques lignes. Merci à Maître Feng, un grand professeur qui sait si bien expliquer en mots et en schémas les notions parfois obscures et contre-intuitives du traitement du signal. Sans ses relectures attentives, son regard critique mais bienveillant, et ses trouvailles pour mettre en valeur des résultats parfois peu engageants, cette thèse ne ressemblerait pas à ce qu'elle est aujourd'hui. Merci à Véronique Aubergé pour m'avoir fait sortir des sentiers battus en me proposant un sujet si impossible à réduire à une seule discipline. Je ne sais pas trop ce à quoi je m'attendais en commençant cette thèse, mais certainement pas à faire autant de rencontres, à discuter avec des gens du théâtre, de l'informatique, des sciences humaines et sociales ou même de la biologie. Que les thèses mono-disciplinaires doivent être ennuyeuses ! Je ferai de mon mieux pour ne pas oublier ce que vous m'avez appris l'un et l'autre. Merci également à Jérôme Maisonnasse et à Germain Lemasson pour m'avoir ouvert les portes du FabMSTIC et accompagnée dans la prise en main de Robair et de tous les outils de prototypage.

Par ailleurs, je remercie mes rapporteurs, Mohamed Chetouani et Jean-François Bonastre, ainsi que les autres membres du jury, Laurent Besacier et Fabienne Martin-Juchat, pour leurs retours très encourageants et leurs questions pertinentes. Grâce à vous, et bien que la soutenance ait eu lieu en semi-présentiel du fait du confinement, j'en garde un excellent souvenir.

Je remercie également Olivier Aycard et Nicolas Marchand pour avoir accepté de constituer mon comité de suivi de thèse et pour leurs retours constructifs ; je sais que ces comités peuvent très mal finir pour les doctorants. Merci également à tous les chercheurs qui ont accepté de prendre un peu de temps pour répondre à mes questions : je pense en particulier à Jean-Sylvain Liénard, qui m'a présenté ses travaux sur la force de voix, et à Nicolas Audibert qui m'a conseillé l'utilisation des modèles linéaires aux effets mixtes. Et merci à Nathalie Vallée pour m'avoir toujours considérée comme un membre de son équipe, malgré notre éloignement géographique.

Une partie importante de cette thèse est consacrée à l'expérimentation, je remercie donc toutes les personnes qui ont accepté de participer à mes expériences, sans oublier les personnels administratifs et techniques, notamment ceux qui se sont occupés de la plateforme Domus pendant mon séjour : Sylvie Humblot-Djamakorzian, Nicolas Bonnefond, Pierre Volcke et Jonathan Bleuzen. Merci également à Christophe Savariaux et Coriandre Vilain, ingénieurs de recherche au GIPSA-lab, pour l'aide qu'ils m'ont apporté en ce qui concerne la calibration acoustique du robot de téléprésence.

Concernant le spectacle Aporia, je remercie Alain Quercia, ainsi que les membres du projet que j'ai pu croiser en régie et pendant les répétitions : Antoine, Benjamin, Clément, Marcel, Émilie... Ce fut un plaisir de travailler avec vous et j'espère que nous aurons à nouveau l'occasion de nous croiser à l'avenir.

En parallèle de mes recherches sur le toucher social, cette thèse m'a donné l'occasion d'encadrer des étudiants. Je remercie donc toutes les personnes qui ont accepté de me confier des heures d'enseignements, en particulier Bertrand Rivet pour m'avoir laissée encadrer des TPs d'électronique, Patrick Reignier pour ses cours de programmation Ada, et Gwenaël Delaval pour m'avoir accueillie au DLST.

Je remercie également mes collègues stagiaires et doctorants : ceux du LIG et du GIPSA-lab bien sûr, que j'ai côtoyés plus ou moins régulièrement au cours de la thèse, mais également tous ceux que j'ai eu l'occasion de rencontrer au cours de diverses formations et séminaires ; sans ces moments d'échanges interdisciplinaires et de convivialité, la thèse perdrait beaucoup de son intérêt et deviendrait très difficile à vivre. En particulier, merci à Yuko, qui a toujours été disponible pour répondre à mes questions, même pendant la rédaction de son manuscrit. Merci également à Romain, Liliya et Natacha avec qui j'ai cohabité à Domus et au cours d'une école d'été à la Sorbonne. Merci aux étudiants Idl et Manintec, et en particuliers à celles qui ont choisi de faire leur stage à Domus : Amélie, Myriam, Émeline et Zoé. Merci aussi à Baptiste. Merci aux habitués de la cafèt' du DPC et aux joueurs de cartes : Alexandre, Andrei, Anne, Firas, Gaël, Omar, Rémy et les autres. Merci à Philip : je te souhaite le meilleur pour la suite de ta thèse.

Enfin, je remercie mes proches pour leur soutien moral. Merci à mes parents pour m'avoir toujours encouragée dans la poursuite de mes études. Merci à mes amis, anciens Houilleux et SICOM, pour m'avoir permis de m'échapper de la thèse. Et surtout : merci Guillaume pour avoir supporté au quotidien mes grands moments de doute et d'enthousiasme. On a réussi !

TABLES DES MATIÈRES

INTRODUCTION

« Un seul être vous manque, et tout est dépeuplé » écrivait Lamartine vers 1820, dans un poème intitulé *L'isolement*. Ce vers résonne encore aujourd'hui à l'oreille de tous ceux qui se retrouvent séparés de ceux qu'ils aiment, que ce soit par la mort, un chagrin d'amour, ou tout simplement l'éloignement. Pourtant, en ce qui concerne les relations à distance, nous vivons à une époque bien différente de celle de Lamartine. **Qu'est-ce qui nous manque, lorsque nous communiquons au téléphone, plutôt qu'en face à face ?** Certainement beaucoup de choses : le toucher, le regard, le fait de partager un espace, ou une activité… À l'inverse, comment expliquer que certaines personnes incapables de se supporter en présentiel se montrent en revanche très cordiales à distance ? **Quels sont les signaux qui passent, et ceux qui ne passent pas, ou sont déformés par la télécommunication au point d'altérer leur sens initial ?** Telles sont les questions qui ont motivées notre recherche, et auxquelles nous allons tenter de répondre au cours de cette thèse.

L'**isolement**, c'est aussi le nom de cette étrange maladie qui touche des personnes âgées parfois très bien entourées, recevant chaque jour la visite d'infirmières, aides à domiciles et femmes de ménages, et qui en épargne d'autres, pourtant plus solitaires, notamment du fait de leur isolement géographique (Mallon 2010). Moins connus en France, les *hikikomoris* incarnent une autre facette de l'isolement relationnel : ces jeunes gens se retirent du monde, en s'enfermant dans leur chambre pendant des mois, voire des années, évitant tout contact IRL[1] (Saito, Angles 2013). Dans les deux cas, on constate qu'il est de plus en plus difficile pour les personnes isolées d'aller vers les autres, comme si elles perdaient progressivement l'habitude et la volonté d'interagir.

Plus largement, le sentiment d'isolement peut apparaître en cas d'hospitalisation, qui vient séparer une personne de ses proches et de son environnement social : en particulier, les enfants ne vivent plus avec leurs parents, et ne peuvent plus suivre une scolarité ordinaire. Dans certains cas, l'isolement fait même partie de l'organisation médicale, lorsque le patient souffre d'une maladie contagieuse, ou au contraire, d'une grave déficience immunitaire ; obligeant les soignants à limiter au maximum le nombre de contacts, et à porter des matériels de protection (masques, gants…) ; ce qui a des conséquences négatives sur la qualité des soins apportés et le bien-être des patients (Abad et al. 2010). Le risque d'isolement est également souvent évoqué dans le cadre du télétravail, pour ses conséquences sur la santé mentale, mais aussi sur l'avancement professionnel des employés, qui pourraient être défavorisés par rapport à leurs collègues plus souvent présents sur place (Tavares 2017).

L'isolement relationnel concerne donc toutes les classes d'âge. C'est un enjeu social, mais également un enjeu de santé publique, puisqu'il s'accompagne d'une dégradation de la santé mentale et physique ; dégradation qui contribue à fragiliser plus encore les personnes isolées. Dans ce contexte, la **robotique de téléprésence** apparaît comme un outil permettant de lutter

[1] « In Real Life » : dans la vraie vie, par opposition aux interactions en ligne, effectuées via le réseau Internet

contre l'isolement social, que ce soit chez les personnes âgées, les jeunes hospitalisés, ou les télétravailleurs. Le principe est de permettre à une personne d'être représentée à distance par un avatar robotique, à travers lequel elle peut percevoir et agir dans un environnement où elle n'est pas. L'objectif est d'aller au-delà du téléphone et de la visioconférence, en parvenant à transmettre non seulement la voix et l'image d'une personne, mais également sa présence.

Plus précisément, notre thèse concerne l'**étude du « toucher social » en téléprésence** ; c'est-à-dire, la manière dont une personne échange des signaux socio-affectifs avec des interlocuteurs distants, signaux qui vont alimenter et donner forme à leur relation. L'expression « toucher social » permet de souligner l'idée que la personne « touchante » et également « touchée » en retour : l'interaction n'est pas un simple échange de messages entre un émetteur et un récepteur, mais sert à construire une relation. En présentiel, le toucher social est à la fois proprioceptif et inter-proprioceptif : non seulement la personne sait exactement ce qu'elle fait, mais elle sait aussi ce que l'autre perçoit de ce qu'elle fait. En téléprésence, ce contrôle est bien plus difficile à réaliser, car il se fait de manière indirecte à travers le robot de téléprésence, ce qui limite les possibilités d'action et de perception du pilote. Notre objectif est donc de permettre au pilote d'un robot de téléprésence d'avoir un toucher social à distance le plus proche possible de celui qu'il aurait en présentiel. Or, ces signaux socio-affectifs que nous manipulons et interprétons tous les jours, sont en réalité très difficiles à saisir.

Le toucher social mobilise principalement trois types de signaux : les signaux vocaux, visuels et tactiles. Nous nous intéresserons principalement au **toucher vocal**, et en particulier à une de ses composantes : la **portée vocale**. C'est à travers sa portée vocale, qu'un locuteur contrôle les personnes qui peuvent l'entendre. En première approximation, la portée vocale dépend uniquement de paramètres physiques : plus une voix est forte, plus sa portée est grande ; et inversement, plus l'environnement est bruyant, plus la portée vocale diminue. Cependant, elle a aussi une dimension sociale importante, puisqu'une mauvaise portée vocale peut nuire à l'interaction ou être perçue négativement : le locuteur s'adapte donc en permanence au contexte acoustique et aux conventions sociales pour ne parler ni trop fort, ni trop doucement. Or, le robot de téléprésence permet de modifier la voix de son utilisateur pour créer des voix qui n'existeraient pas sans technologie. En particulier, une voix douce peut-être amplifiée pour devenir audible à grande distance ; ou une voix forte atténuée jusqu'au niveau d'un chuchotement. Dès lors, comment définir ce qui serait une bonne portée vocale en téléprésence ? S'agit-il uniquement d'une question d'intensité, ou faut-il également tenir compte des socio-affects exprimés, ou encore de la position du robot dans l'espace ? C'est ce que nous chercherons à définir à travers de nouvelles expériences.

Nous aborderons la question de la **portée vocale** à travers deux approches distinctes. D'une part, nous nous inspirerons des études sur la localisation spatiale en psychoacoustique pour réaliser une étude sur la perception de la portée vocale. D'autre part, nous chercherons à quantifier l'effet Lombard en robotique de téléprésence, afin de déterminer si la portée vocale du pilote du robot peut être altérée en présence de bruit.

Notre thèse s'inscrit dans une recherche au long cours visant à comprendre la nature du **lien social** à travers l'étude des signaux socio-affectifs. Cette recherche s'est d'abord appuyée sur

des travaux appliqués à la synthèse vocale ou à l'apprentissage des langues (Aubergé 1991), (Rilliard 2000), (Loyau 2007), (Shochi 2008), (Audibert 2008), (Vanpé 2011), (Mac 2012), (Lu 2015), puis plus récemment à la robotique (Sasa 2018). Elle repose sur une approche holistique de la parole, qui ne se limite pas à la voix seule, et est étudiée comme le fruit d'un contexte social, biologique et environnemental. L'humain est également abordé comme un individu autonome, mais pas indépendant, car lié socialement aux autres[2]. La méthode scientifique qui en découle repose sur une expérimentation « écologique »[3], au sens où les sujets de l'expérience sont placés dans des conditions les plus proches possibles de celles dans lesquelles les technologies étudiées seront amenées à être utilisées.

La thèse est découpée en six chapitres. Le premier sera consacré aux **robots de téléprésence**. Nous commencerons par définir la téléprésence et les notions qui lui sont associés ; puis à travers l'exemple de plusieurs prototypes avancés développés pour la recherche, nous présenterons les prérequis technologiques nécessaires. Ensuite, nous nous intéresserons aux robots de téléprésence « grand public » tels qu'ils existent à l'heure actuelle, afin de mettre en avant leurs caractéristiques et leurs défauts. En particulier, nous verrons que les innovations apportées à ces robots concernent principalement la navigation en toute sécurité, au détriment peut-être d'une réflexion sur le toucher social : ainsi, la question posée par les roboticiens est d'abord « Comment éviter les obstacles ? » plutôt que « Comment agir à distance dans un environnement social ? ». De plus, contrairement aux prototypes plus avancés, ces robots « grand public » permettent une téléprésence ubiquïte : il ne s'agit pas de transporter la personne d'un lieu à un autre, mais de lui permettre d'être à deux endroits à la fois. Nous présenterons également des technologies pour l'immersion acoustique à distance.

Dans le second chapitre, nous nous intéresserons au **toucher vocal**, et aux outils permettant de le décrire. Nous verrons que ce toucher vocal est indissociable du contexte physique et social de l'interaction. Nous proposerons de tracer les contours de la portée vocale ; puis nous présenterons en détails une série d'articles qui montrent que lorsqu'un auditeur perçoit la distance qui le sépare d'un locuteur, il est en partie influencé par sa portée vocale.

Ensuite, au chapitre 3, nous présenterons les **grands principes méthodologiques** qui ont guidé nos travaux de recherche. Leur objectif est de pouvoir étudier la parole et développer des technologies dans des conditions de laboratoire les plus proches possibles des conditions écologiques ; c'est-à-dire les conditions d'utilisation réelles de ces technologies.

Les chapitres suivants seront consacrés aux **travaux réalisés** pendant la thèse. Au chapitre 4, nous présenterons une série d'études consacrées à la perception de l'espace physique et social : l'objectif sera de déterminer si ces deux domaines sont bien distincts, ou, au contraire, s'ils s'enchevêtrent au point qu'une variation de l'un engendre une variation dans la perception de l'autre. Par exemple, est-ce qu'une voix douce est perçue plus proche qu'une voix dure ? Et inversement, est-ce que la distance modifie la perception des socio-affects ? Autrement dit,

[2] L'approche opposée serait un idéalisme (« on peut séparer le corps de l'esprit ») réductionniste (« le tout peut s'expliquer par la somme de ses parties »).
[3] Avant de désigner un mouvement politique, l'écologie est d'abord une discipline scientifique, consacrée à l'étude des êtres vivants et de leurs interactions dans leur milieu.

nous cherchons par ces expériences à identifier les contraintes physiques qui pèsent sur le toucher vocal, et qui sont susceptibles d'être modifiées en téléprésence.

Puis, nous nous intéresserons au chapitre 5 à un cas particulier de la robotique de téléprésence ubiquïte : Que se passe-t-il lorsque le pilote du robot entend un bruit, que ses interlocuteurs ne perçoivent pas ? Est-ce que le fait d'être présent à deux endroits à la fois suffit à altérer le toucher vocal ?

Enfin, le chapitre 6 sera consacré à notre participation à Aporia, un spectacle Arts-Sciences, sur le thème de l'altérité. Il s'agit d'une adaptation d'un texte de Benard-Marie Koltès, au cours de laquelle un acteur modifie son toucher vocal pour incarner les différents personnages de la pièce. Pour ce faire, il s'appuie sur un outil numérique qui transforme sa voix en temps-réel, et la diffuse via des haut-parleurs.

Chapitre 1 :
ROBOTS DE TÉLÉPRÉSENCE ET IMMERSION SOCIALE À DISTANCE

Dans le roman « La terre bleue de nos souvenirs », (Reynolds 2015) imagine un futur où l'humanité a colonisé le système solaire. Les robots de téléprésence font désormais partie du quotidien d'un certain nombre de personnes. À l'aide de microprocesseurs implantés dans leurs cerveaux, capables de manipuler leurs sensations, elles peuvent apparaître à leurs interlocuteurs sous la forme de « chimères » virtuelles, piloter un robot à distance, ou encore prendre possession du corps d'un individu consentant, comme s'il était le leur. Les hommes et femmes d'affaires les plus fortunées disposent même d'androïdes leur ressemblant trait pour trait : l'illusion est si parfaite, qu'on a l'impression de parler avec un véritable être humain. En outre, ces robots sont programmés pour pouvoir fonctionner en mode semi-autonome : ils sont capables d'imiter le comportement de leur propriétaire pour le représenter au cours de soirées de gala ennuyeuses, ou de faire acte de présence à des réunions interplanétaires, lorsque les délais de transmissions sont trop importants pour permettre une interaction directe.

Pour l'instant, ce type de téléprésence relève de la science-fiction. Pourtant, l'univers décrit par (Reynolds 2015) illustre parfaitement les principales thématiques de la recherche actuelle en robotique de téléprésence, comme nous le verrons dans ce chapitre. Nous présenterons également un catalogue des robots de téléprésence disponibles dans le commerce, pour en extraire les principales caractéristiques. Puis, nous nous intéresserons aux limites techniques de ces robots, et à leurs conséquences sur l'interaction. En particulier, nous verrons que la question du « toucher vocal » n'est que rarement abordée ; bien qu'il existe aujourd'hui des technologies grand public très accessible permettant une immersion acoustique à distance.

1.1 Des robots de téléprésence pour augmenter les capacités de télécommunication…

Commençons par quelques définitions. Un **robot de téléprésence** est un système de télécommunication, à l'instar du téléphone et de la visioconférence. Si le téléphone transporte uniquement la voix d'une personne d'un lieu à un autre, et la visioconférence à la fois la voix et l'image, l'objectif du robot de téléprésence est d'aller encore plus loin, en transportant également la sensation de présence.

Cette sensation de **présence** est très subjective et complexe à définir, d'autant qu'elle a été étudiée par plusieurs disciplines, qui n'ont pas toujours la même terminologie, ou le même objet d'études (Lee 2004). En particulier, certains concepts ont été pensés pour la réalité virtuelle, et ne sont donc pas toujours pertinents en robotique de téléprésence : par exemple, parler de « réalisme », n'a pas de sens lorsqu'on interagit avec un environnement réel.

Initialement, l'expression téléprésence apparaît pour la première fois dans le manifeste de (Minsky 1980), qui plaide pour le développement de machines capables de reproduire à l'identique les gestes de leur opérateur. Minsky pense en particulier aux professions dangereuses, tels que les mineurs ou les travailleurs du nucléaire, qui devraient pouvoir réaliser leurs tâches à distance en toute sécurité. Cette forme de présence à distance est parfois qualifiée de « *presence as transportation* » (Nowak 2001), ou de **présence physique** (Lee 2004).

Selon la classification proposée par (Lee 2004), il existe deux autres formes de présence : la présence sociale et la présence de soi (*self-presence*). La **présence sociale** désigne le fait de se sentir au contact de personnes réelles. En robotique de téléprésence, elle concerne en particulier la manière dont le robot est perçu par les interlocuteurs. Enfin, la **présence de soi** désigne le degré d'identification de l'utilisateur à son avatar (virtuel ou robotique).

La notion d'**immersion** est également souvent évoquée, et tout aussi complexe à définir. Ainsi, certains auteurs utilisent le mot immersion pour parler d'une qualité objective : plus un système permet à une personne de se sentir immergée sensoriellement dans un autre environnement, plus il est immersif. Pourtant, la sensation éprouvée finalement par l'utilisateur reste éminemment subjective : bien qu'elle soit très certainement influencée par la qualité immersive du média utilisé, l'immersion désigne aussi un état perceptif, qui peut être provoqué de différentes manières. Par exemple, un lecteur peut être immergé dans sa lecture : c'est-à-dire que son attention est entièrement focalisée sur le récit. De même, un joueur de Tetris peut être absorbé par son jeu. (Nilsson et al. 2016) distinguent ainsi trois formes d'immersion : l'immersion technologique (*system immersion*), l'immersion narrative (*narrative immersion*) et l'immersion basée sur le challenge (*challenge-based immersion*). En robotique de téléprésence, c'est la première forme d'immersion qui nous intéresse.

À défaut de pouvoir transporter instantanément une personne d'un endroit à l'autre, la robotique de téléprésence a donc initialement pour objectif de créer une immersion technologique, permettant de susciter trois sentiments de présence : présence physique, présence sociale et présence de soi. Nous regrouperons les recherches qui vont dans cette direction dans une grande

famille : celle du ***Beaming***, qui est une traduction possible en anglais du mot téléportation[4]. S'il existe à l'heure actuelle plusieurs prototypes de *Beaming*, comme nous le verrons dans la suite, ceux-ci restent encore à l'état d'études.

À l'inverse, les robots de téléprésence disponibles actuellement dans le commerce proposent plutôt d'une forme de téléprésence **ubiquïte** : leur utilisateur n'est pas séparé de son environnement, mais présent à deux endroits à la fois. Par convention, l'environnement du robot est généralement qualifié d'espace « local », puisque c'est là que se déroule l'interaction, tandis que l'environnement du pilote est qualifié d'espace « distant » (cf. Figure 1). Dans le cadre de cette thèse, c'est cette forme de téléprésence, et les problématiques qui en découlent, qui nous intéressent particulièrement.

Figure 1 : Téléprésence ubiquïte

Dans cette première partie, nous présenterons à la fois les recherches concernant le Beaming, puis nous ferons un état des lieux de la robotique de téléprésence grand public.

1.1.1 Prototypes de Beaming en laboratoire

Nous allons d'abord présenter une sélection de travaux qui se rattachent au *Beaming*. Chacun concerne un aspect de la téléprésence évoqué précédemment : présence physique, présence sociale ou présence de soi. Il ne s'agit pas ici d'un état des lieux exhaustif, mais d'une entrée en matière, qui présente rapidement ce qui existe aujourd'hui dans les laboratoires de robotique.

1.1.1.1 Des sensations et des mouvements reproduits le plus fidèlement possible

Pour que le pilote du robot de téléprésence se sente physiquement présent à distance, il doit pouvoir percevoir et agir dans l'environnement local comme s'il y était. La qualité de cette immersion repose sur plusieurs variables technologiques, dont (Steuer 1992) propose une classification (cf. Figure 2).

[4] En particulier, « Beam me up, Scotty. » est une phrase emblématique de la série Star Trek, par laquelle le capitaine Kirk ordonne à son ingénieur en chef d'activer le téléporteur qui le ramène au vaisseau spatial.

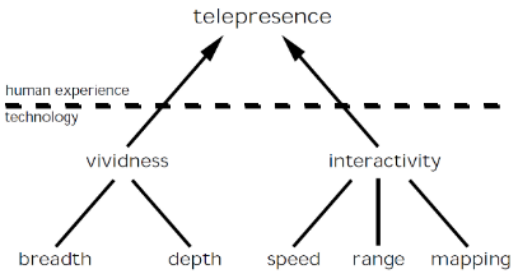


Figure 2 : Variables technologiques influençant la téléprésence (Steuer 1992)

Selon cette classification, il existe deux variables principales, qui sont :

- la **richesse** de l'environnement (*vividness*), déclinable en deux catégories : l'ampleur sensorielle (*breadth :* nombre de sens mobilisés) et la profondeur (*depth* : résolution de chaque canal sensoriel). Une liste des sens à mobiliser est proposée par (Mair 2007) : elle inclut les sens visuel, auditif, haptique (toucher et proprioception), mais également olfactif et vestibulaire (perception de la position et des mouvements de la tête liée à l'oreille interne et nécessaire à l'équilibre). Seul le sens du goût n'est pas évoqué par Mair.

- l'**interactivité** (possibilité de l'opérateur de modifier son environnement en temps réel), déclinable en trois catégories principales : la vitesse de réponse du système aux actions de l'opérateur (*speed*), la gamme des actions possibles (*range*) et la modélisation (*mapping*) qui associe chaque contrôle à un effet sur le monde virtuel de manière à rendre naturel et prévisible le résultat des actions de l'opérateur.

Cette classification, prévue initialement pour la réalité virtuelle, est tout aussi pertinente en robotique de téléprésence[5].

Un exemple actuel qui illustre cet aspect de la téléprésence est fourni en Figure 3. Il s'agit d'un arrêt sur image d'une vidéo de démonstration du projet Sombrero (Gipsa-lab). On y voit un robot anthropomorphe asservi aux mouvements de son pilote : lorsque le pilote tourne la tête, le robot tourne également. Le pilote est équipé d'un casque de réalité virtuelle, qui lui permet de voir à travers les « yeux » du robot. Un capteur « Leap Motion » est posé sur la table, pour enregistrer les gestes des mains du pilote et les reproduire sur le robot.

[5] Cependant, la robotique de téléprésence pose une difficulté supplémentaire : l'environnement dans lequel le pilote interagit est un environnement physique réel, pas une simulation. Les sens du pilote doivent donc être reproduits le plus fidèlement possible par rapport à la réalité ; tandis qu'une simulation peut se permettre de plus de libertés. Par exemple, en réalité virtuelle, il est possible de simuler une allée fleurie et odorante de façon très réaliste, en choisissant de représenter des fleurs dont le parfum est facilement accessible. Il serait très difficile d'atteindre une sensation équivalente en robotique téléprésence, car il faudrait pouvoir recréer n'importe quelle odeur.

Figure 3 : Téléopération immersive du robot ROMEO
(Extrait d'une vidéo de démonstration du projet Sombrero,
http://www.gipsa-lab.fr/projet/SOMBRERO/videos.html)

On peut également citer les recherches qui concernent le sens du toucher, notamment en téléchirurgie (Okamura 2004). En effet, pour pouvoir opérer à distance, le chirurgien a besoin d'un retour concernant la force du robot (*haptic feedback*). Ce retour peut être fourni de façon directe (par exemple, la résistance mécanique de l'appareil évolue en fonction des mouvements du chirurgien); ou de manière indirecte, en traduisant la force exercée par le robot en informations visuelles ou acoustiques.

1.1.1.2 Une apparence qui imite l'humain

Les robots utilisés pour réaliser un *Beaming* sont nécessairement anthropomorphes, puisqu'ils sont sensés pouvoir reproduire le plus fidèlement possible les gestes d'un être humain. Certains vont encore plus loin dans l'imitation : on peut citer en particulier, les Geminoids, développés initialement par l'équipe du japonais Hiroshi Ishiguro. Ces androïdes sont conçus pour imiter le plus fidèlement possible l'apparence de leur propriétaire (cf. Figure 4). Ishiguro fait en effet l'hypothèse que plus les robots ressembleront à des humains, plus il sera facile d'interagir avec eux (présence sociale). Le chercheur a choisi de copier l'apparence de personnes existantes (lui-même et sa fille), afin de pouvoir comparer aisément les interactions humain-robot et humain-humain. Son objectif est de saisir ce qui fait la présence humaine, à travers des expériences de téléopération du clone robotique par son propriétaire.

Figure 4 : Photos de Risa Ishiguro, Hiroshi Ishiguro et Henrik Scharfe, accompagnés de leurs Geminoids.
(http://www.geminoid.jp/en/index.html)

Quoique d'une ressemblance saisissante dans leur forme et leur texture, ces robots restent imparfaits, car ils échouent à reproduire les détails fins mais essentiels de la gestualité et des expressions. Ils peuvent ainsi provoquer une réaction de rejet chez leur interlocuteur. Ce

phénomène est connu sous le nom de « vallée de l'étrange » (*bukimi no tani*, *uncanny valley*). Il a été théorisé dès 1970 par le roboticien Mori Masahiro, qui rejette la croyance selon laquelle plus l'apparence des robots se rapproche de celle des humains, mieux ils seront perçus. Au contraire, avant d'atteindre son « pic », en devenant indiscernable d'une personne humaine, le robot anthropomorphe passerait par une vallée de l'étrange, dans laquelle ses petites imperfections seraient perçues comme anormales et dérangeantes.

1.1.1.3 Une projection dans le corps du robot

Un des enjeux principal du *Beaming* est de donner au téléopérateur la sensation que le corps du robot est le sien (présence de soi). Or, il existe un domaine de recherche dédié aux conditions dans lesquelles il est possible de créer cette illusion de propriété corporelle (*body ownership illusion*). On peut citer notamment l'expérience de (Petkova, Ehrsson 2008), qui consiste à « échanger » le corps d'un sujet avec celui d'un mannequin du sexe opposé. Le mannequin, debout, est équipé d'une caméra qui filme vers le bas. Le sujet, équipé d'un casque de réalité virtuelle, est placé dans la même position que le mannequin et reçoit le flux vidéo filmé par la caméra du mannequin : à la place de son corps, c'est donc le corps du mannequin qu'il voit. Après une série de stimulations tactiles appliquées simultanément au sujet et au mannequin, le sujet a une réaction de recul lorsqu'un couteau est approché du ventre du mannequin.

Cette méthodologie a été reprise un collectif Arts-Sciences : le « Be Another Lab », qui explore la relation entre identité et empathie, en utilisant la réalité virtuelle pour permettre à des personnes de voir à travers les yeux d'un autre. Dans leur expérience, l'échange de corps n'a pas lieu avec un mannequin, mais entre deux personnes, qui doivent bouger ensembles pour maintenir l'illusion en place. D'après les témoignages de ceux qui l'ont vécue, l'expérience peut être très émouvante, en particulier quand l'échange a lieu avec un inconnu.

1.1.1.4 Des robots potentiellement autonomes ou semi-autonomes

Lorsqu'on parle de téléopération à distance, la question de l'automatisation peut rapidement apparaître. Ainsi, (Minsky 1980) l'évoque dès son manifeste pour la téléprésence : avec le développement de l'intelligence artificielle, les téléopérateurs humains finiraient par déléguer leurs tâches aux robots, pour devenir leurs superviseurs. (Nishio et al. 2007) ont également prévu la possibilité que les Geminoids puissent servir de substitut à la présence de leur propriétaire : ils agiraient de façon semi-autonome, et seraient téléopérés uniquement dans des situations où le contrôle humain est nécessaire.

Le projet Sombrero est un exemple concret où l'objectif du *Beaming* est d'« apprendre » à un robot à imiter les comportements humains, en lui fournissant des exemples d'interactions humain-humain adaptés à ses capacités motrices. Ainsi les robots utilisés dans le cadre de ce projet ne sont pas initialement des robots de téléprésence. Le robot Romeo est le robot majordome de l'entreprise SoftBank Robotics, conçu pour aider des personnes âgées en perte d'autonomie. De même, le robot Nina est un iCub, plateforme de robotique humanoïde développée dans le cadre du projet européen RobotCub pour le développement d'algorithmes d'intelligence artificielle.

1.1.2 La robotique de téléprésence grand public

Les robots de téléprésence anthropomorphes présentés précédemment restent très minoritaires, et cantonnés quasi exclusivement aux laboratoires de recherche. Il n'est pas possible pour un particulier de commander un Geminoid, ou un Romeo. En revanche, il existe un certain nombre de robots de téléprésence achetables dans le commerce, à la production industrialisée. La Figure 6 en présente un échantillon quasi exhaustif, que nous allons analyser pour en extraire les principales caractéristiques.

1.1.2.1 Apparence standard : un écran à base mobile

La plupart de ces robots reposent sur le même principe : un écran-tête équipé d'une caméra qui affiche l'image du pilote, et une base mobile. Ils permettent ainsi à leur utilisateur de voir, d'être vu, et de se déplacer dans l'environnement local. Seule la taille de l'écran et la forme de la base mobile varient d'un constructeur à l'autre. Pour la moitié d'entre eux, l'écran est placé à l'horizontale, tel un écran de télévision classique. Pour l'autre moitié, au contraire, l'écran est placé à la verticale, ce qui permet d'afficher uniquement le visage du pilote. En général, l'écran constitue une grande partie du corps du robot. Il est ainsi mis en avant par rapport à la base mobile, dont le but est avant tout utilitaire : elle doit supporter l'écran, permettre au robot de bouger, et de se recharger. Quelques robots, tels que le Jazz, le Synergy Swan ou le QB, ont un aspect plus robotique, du fait d'un écran plus discret. En particulier, les caméras et le haut-parleur du QB sont placés de manière à évoquer un visage.

On compte tout de même une exception notable à ce modèle « écran-base mobile » : le robot Collaborate i/o, qui prend la forme d'un simple œil haute-résolution, monté sur un bras articulé. Il répond à un besoin très spécifique : permettre à des ingénieurs de diagnostiquer des problèmes techniques à distance, en leur fournissant la meilleure vision possible. Bien que présenté par ses concepteurs comme un robot de téléprésence, son principe est assez différent, puisqu'il nécessite la collaboration entre deux personnes. La première, présente dans l'environnement local, doit positionner grossièrement l'œil au-dessus de l'objet à étudier (ex : circuit électronique). La seconde, téléprésente, doit ensuite régler finement la caméra, afin d'obtenir l'image la plus nette possible.

Notons également que la palette de couleurs de ces robots de téléprésence est assez limitée, puisqu'on trouve principalement du blanc, du noir, ou du gris. Seul le robot suédois Giraff détonne avec son bleu électrique (il existe également en version blanche).

1.1.2.2 Des robots à mobilité variable

Par définition, un robot de téléprésence est mobile, contrairement à un système de visioconférence. Cependant, tous n'ont pas les mêmes degrés de liberté. Un certain nombre sont de taille ajustable, pour permettre à leur pilote de s'adapter à la posture de ses interlocuteurs (assis ou debout). Ils sont accompagnés sur la figure ci-dessus d'une flèche bleue verticale qui représente la hauteur maximale et minimale du robot. Les robots accompagnés d'une flèche verte courbe sont ceux dont la tête est inclinable.

Si la plupart sont équipés de roues, il existe également des robots de téléprésence stationnaires, par exemple les robots Collaborate i/o, Kubi, Tabletop TeleMe, Swivl et SelfieBot. Bien qu'ils ne permettent pas à leur utilisateur de se déplacer dans l'environnement local, ils restent plus mobiles que des systèmes de visioconférence classique, puisqu'ils peuvent pivoter sur eux-mêmes.

1.1.2.3 Sens du toucher absent

Sauf exception, les robots de téléprésence ne permettent pas d'attraper des objets. Parmi les exemples cités, seul le robot Peoplebot possède une pince articulée, capable de se déplacer sur un axe vertical et de saisir de petits objets (ex : bouteille plastique). Cependant, bien que pouvant être utilisé en téléprésence, ce robot est d'abord une plateforme conçue pour la recherche en robotique.

1.1.2.4 Un prix qui dépend du type de capteurs utilisés

Le prix des robots de téléprésence est extrêmement variable : de quelques centaines d'euros, à plusieurs milliers. Cette variabilité s'explique principalement par le type de capteurs utilisés. Ainsi, les modèles les plus bas de gamme ne sont que des supports mobiles, sur lequel l'utilisateur doit placer sa propre tablette/smartphone. Pour les modèles plus sophistiqués, l'écran est intégré au robot, ainsi que des microphones et une ou plusieurs caméras de bonne qualité. Les plus onéreux, tels que le PeopleBot ou le Vita, sont équipés de lidars et d'algorithmes d'aide à la navigation.

1.1.2.5 Trois secteurs d'activité : le télétravail, la santé et l'éducation

Dans leur revue, (Kristoffersson et al. 2013) identifient quatre environnements dans lesquels les robots de téléprésence sont principalement déployés : les environnements de bureaux, les environnements de soin, les lieux de vie des personnes âgées, et les environnements scolaires. Autrement dit, trois secteurs d'activité sont particulièrement visés : le télétravail, la santé, et l'éducation.

En entreprise, le robot de téléprésence doit permettre aux salariés de mieux communiquer à distance. Par exemple, les fabricants du robot Double mettent en avant l'importance des discussions informelles quotidiennes, peu adaptées à la visioconférence classique qui doit être planifiée à l'avance. Au contraire, il est possible avec un robot de téléprésence de circuler librement dans les bureaux, et d'être physiquement présent à son poste pendant les heures de travail. Le télétravailleur ne risquerait donc plus d'être isolé par son équipe, puisqu'il conserve une présence physique au sein de l'entreprise.

Du côté de la santé, on peut citer notamment Vita, robot de télémédecine, conçu pour permettre à un médecin d'effectuer des consultations à distance. Il est équipé d'un stéthoscope, qui lui permet d'écouter le cœur du patient avec l'aide d'un infirmier présent dans l'environnement local. Le robot Giraff, lui, a été développé spécifiquement pour le soin aux seniors (Cesta et al. 2016). Il est déployé chez des personnes âgées, afin de permettre à leurs médecins et à leurs proches de leur rendre visite régulièrement.

Dans le domaine de l'éducation, il s'agit d'enseignement à distance ; le robot pouvant être piloté soit par un élève, soit par un enseignant. En particulier, les robots QB et Beam+ ont été utilisés au cours de l'expérimentation « Robot Lycéen », pour permettre à des élèves hospitalisés de la région Rhône-Alpes de continuer à suivre leurs cours à distance (Coureau-Falquerho et al. 2017). On compte également plusieurs études concernant l'utilisation de robots de téléprésence par des professeurs d'anglais : par exemple, celles des coréens (Oh-Hun Kwon et al. 2010) ou des japonais (Tanaka et al. 2014). En effet, il existe en Asie une forte demande d'enseignants de langue anglaise, pour un faible nombre de locuteurs dont c'est la langue maternelle.

1.1.2.6 Des robots de service intégrant des fonctionnalités de vidéoconférence

Par ailleurs, notons que plusieurs robots qui ne sont pas initialement conçus pour la téléprésence, intègrent des fonctionnalités de visioconférence. La Figure 5 en présente un petit échantillon. Il peut s'agir de robots de présentation, conçus pour accueillir et conseiller les clients d'une entreprise ; ou de robots de surveillance, conçus pour patrouiller dans une zone. En principe, un opérateur humain supervise une flotte de robots, et peut prendre la main sur un d'entre eux, par exemple lorsqu'un client pose une question à laquelle le robot ne peut pas répondre. De même, les robots majordomes ou compagnons peuvent être utilisés comme outils de télécommunication, puisqu'ils sont déjà équipés de caméra, écran, microphone et haut-parleur. Ces robots sont cependant plus onéreux que les simples robots de téléprésence, puisque pour pouvoir fonctionner de façon autonome, ils intègrent des algorithmes et des capteurs plus sophistiqués. Leur apparence est également différente de celles des robots de téléprésence : généralement plus massive pour les robots de présentation, et plus anthropomorphe pour les robots majordomes.

Figure 5 : Quelques robots de service

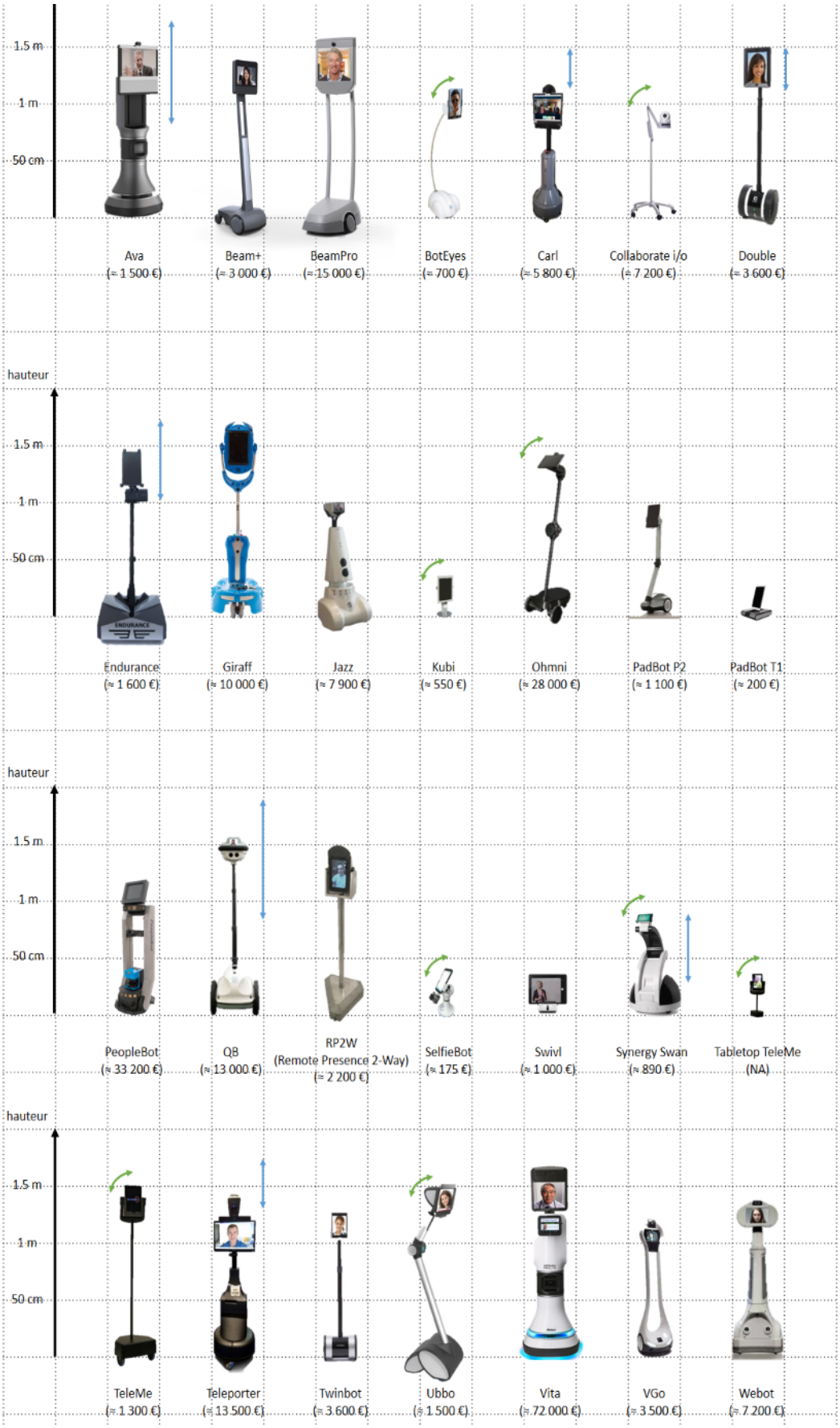


Figure 6 : Robots de téléprésence disponibles dans le commerce
(prix indicatifs obtenus à partir des sites d'achat en ligne telepresencerobots.com et robotshop.com)

1.1.3 Possibilités offertes par les robots de téléprésence actuels

Bien qu'ils ne permettent pas de réaliser un véritable *Beaming*, les robots de téléprésence disponibles à l'heure actuelle constituent déjà une avancée notable par rapport aux systèmes de visioconférences classiques, comme le montrent plusieurs cas d'usages, notamment (Neustaedter et al. 2016 ; Yang et al. 2018).

1.1.3.1 Une présence accrue

En donnant corps à leurs pilotes, les robots de téléprésence leur confèrent une autonomie et une présence physique et sociale bien supérieure à celle qu'ils auraient en simple visioconférence. L'interlocuteur distant n'est plus entièrement dépendant des personnes présentes dans l'environnement local : il peut par exemple déplacer le robot dans la pièce, ou modifier l'angle de la caméra s'il le souhaite. Par ailleurs, ces robots de téléprésence sont sensés permettre à une personne non seulement de communiquer à distance, mais surtout de participer à des activités sociales. Ainsi, dans le cadre du télétravail, un des cas d'usages mis en avant du robot de téléprésence est de permettre à un employé de continuer à avoir des discussions informelles avec ses collègues, à la pause-café par exemple. De même, un enfant qui suivrait sa scolarité en téléprésence devrait avoir la possibilité de participer aux récréations, pour pouvoir jouer avec ses camarades de classe.

1.1.3.2 Une perception augmentée de l'environnement local

Le robot permet également d'augmenter les capacités perceptives de son pilote. Ainsi, le robot Collaborate i/o est équipé de lentilles optiques et de leds qui permettent à son pilote de voir à distance, encore plus précisément que s'il était sur place. Couplé à une base de données et un système de réalité augmentée, le robot de téléprésence pourrait permettre à son pilote d'avoir accès à plus d'informations sur l'environnement local que s'il y était présent en chair et en os : à travers son interface, il pourrait connaître la température de la pièce, avoir accès au plan du bâtiment, ou encore voir le nom de ses interlocuteurs apparaître en surimpression. En ce qui concerne l'audition, plusieurs robots de téléprésence intègrent des technologies d'annulation de bruit, pour rehausser la voix des personnes à proximité du robot. (Liu et al. 2015) vont jusqu'à proposer un système permettant de contrôler son environnement sonore en modifiant virtuellement le volume de chaque personne présente dans la pièce. Ainsi, le pilote pourrait « couper » les voix qui ne l'intéressent pas, et créer son propre environnement sonore virtuel.

1.1.3.3 Une apparence choisie

Par ailleurs, on peut imaginer que le pilote ait la possibilité de choisir l'apparence du robot qui le représente. Le besoin de personnalisation du robot afin de rappeler l'identité de son propriétaire a notamment été soulevé au cours de tests d'usage, en particulier dans le cas où plusieurs robots du même type sont présents (Neustaedter et al. 2016). Cette personnalisation peut passer notamment par le choix de vêtements pour habiller le robot (Lu et al. 2011 ; Saadatian et al. 2013). Si le but peut être de rapprocher l'apparence du robot de celle de son utilisateur, la personnalisation peut tout aussi bien servir à un travestissement. Ainsi, un adepte du bodybuilding à la carrure imposante pourrait choisir d'incarner un petit robot à l'aspect fragile et « mignon » plutôt qu'un *Terminator*. Il est également envisageable d'utiliser des filtres

animés pour modifier une vidéo en temps-réel : le pilote pourrait choisir de porter un masque, ou de modifier la forme de son visage.

La voix du pilote est également concernée par ces augmentations technologiques, puisqu'il devient envisageable de modifier artificiellement ses caractéristiques, soit pour changer son identité, soit pour augmenter son intelligibilité. Cette possibilité de transformer sa voix permet d'imaginer de nouvelles formes d'expression, de la même manière que l'utilisation du microphone a permis l'émergence de nouvelles formes musicales[6]. Le pilote du robot pourrait choisir de rendre sa voix plus aiguë, plus forte, ou plus résonnante pour appuyer ce qu'il dit. Les personnes insatisfaites du son de leur voix pourraient s'en créer une nouvelle, plus mélodieuse, ou plus rauque par exemple.

1.1.3.4 Un facilitateur d'interactions

À l'instar du téléphone, le robot de téléprésence augmente les capacités de communication de son utilisateur, en lui permettant de communiquer à distance. Dans certains cas, il est même présenté comme un outil thérapeutique, qui permettrait d'améliorer les relations sociales, malgré la distance. Un salarié qui télétravaille ne risquerait plus de s'isoler de ses collaborateurs ; un médecin pourrait rendre régulièrement visite à ses patients pour prendre de leurs nouvelles… Moins intimidants qu'un fauteuil roulant lourdement équipé, les robots de téléprésence permettraient également de donner une mobilité nouvelle à des personnes immobilisées du fait de leur traitement ou d'un handicap physique lourd. Ainsi, (Leeb et al. 2015) se sont intéressés au contrôle d'un robot de téléprésence à l'aide d'une interface cerveau-machine basée sur les signaux EEG[7] destinée à des personnes qui ne pourraient pas utiliser un pilotage manuel.

1.1.4 Résumé

Un robot de téléprésence est un système de télécommunication, conçu pour transmettre la présence de ses utilisateurs. Cette présence peut être simulée par la technologie, en particulier en utilisant des systèmes de diffusion immersifs, capables de s'adapter en temps-réel aux mouvements du pilote. L'enjeu initial est de permettre à une personne de se « téléporter » d'un endroit à l'autre, afin de pouvoir agir dans un environnement distant comme si elle y était. Il s'agit donc dans un premier temps de parvenir à reproduire à l'identique les sensations et les gestes du pilote. Dans un second temps, le robot de téléprésence pourrait augmenter son utilisateur, en lui permettant de percevoir et d'agir dans l'environnement « mieux » que s'il y était présent. Il existe un écart important entre les prototypes de *Beaming* développés en laboratoires, et les robots de téléprésence grand public. Ces derniers se limitent encore pour la plupart à un simple écran, monté sur une base mobile. La ressemblance entre les différents modèles est d'ailleurs frappante, bien que la cible commerciale varie d'un constructeur à l'autre.

---

[6] Jusque dans les années 30, les chanteurs populaires étaient obligés de chanter fort, à la manière des chanteurs d'opéra, pour pouvoir être entendus de leur public ; mais l'utilisation du microphone et de l'amplification a permis l'émergence du style des *crooners*, appréciés pour le timbre de voix sensuel et intime (Chermayeff, Le Goff 2017).

[7] EEG (Électroencéphalogramme) : mesure de l'activité électrique du cerveau à l'aide d'électrodes placées sur le cuir chevelu

1.2 … mais sources d'artefacts relationnels

Si les robots de téléprésence commencent à être disponibles sur le marché à des prix abordables, leur usage reste limité, et n'a pas encore remplacé les autres systèmes de télécommunication. En effet, bien qu'ils offrent de nouvelles possibilités intéressantes, ils souffrent d'un certain nombre de défauts, qui les rendent complexes à manipuler.

1.2.1 Limites techniques de la télétransmission

Les systèmes de télécommunication sont des systèmes physiques : ils ne peuvent pas transmettre une quantité d'information infinie de façon instantanée. Les robots de téléprésence sont donc soumis à un certain nombre de contraintes, qui dépendent du réseau utilisé pour transmettre l'information entre l'espace distant et l'espace local.

1.2.1.1 Utilisation du réseau Internet pour le transfert de données

Pour pouvoir piloter à distance un robot de téléprésence, il faut parvenir à échanger un certain nombre d'informations entre l'ordinateur du pilote et le robot : à la fois les commandes envoyées au robot, mais aussi les flux audio et vidéo permettant la communication entre le pilote et ses interlocuteurs.

Sauf cas particulier, c'est le réseau Internet qui est utilisé pour réaliser ces échanges. Or ce réseau n'est pas très adapté aux communications en temps-réel. En effet, pour acheminer l'information d'un point à un autre du réseau, il repose sur un système de commutation par paquets : les données à échanger sont découpées en paquets d'informations, étiquetés en fonction de leur contenu et de leur destination. Chaque paquet est ensuite acheminé à travers le réseau, parfois à travers des chemins différents en fonction des chemins disponibles à chaque instant. Une fois arrivés à destination, les paquets doivent être réassemblés dans le bon ordre. Au contraire, les premiers réseaux téléphoniques reposaient sur une commutation par circuit : une fois le chemin de transmission établi, celui-ci était maintenu pour toute la durée de la communication. Bien plus simple à mettre en œuvre, cette commutation par circuit a le désavantage de monopoliser le réseau : tant que l'appel téléphonique n'est pas terminé, il est impossible de transmettre d'autres informations sur la ligne occupée (cf. Figure 7).

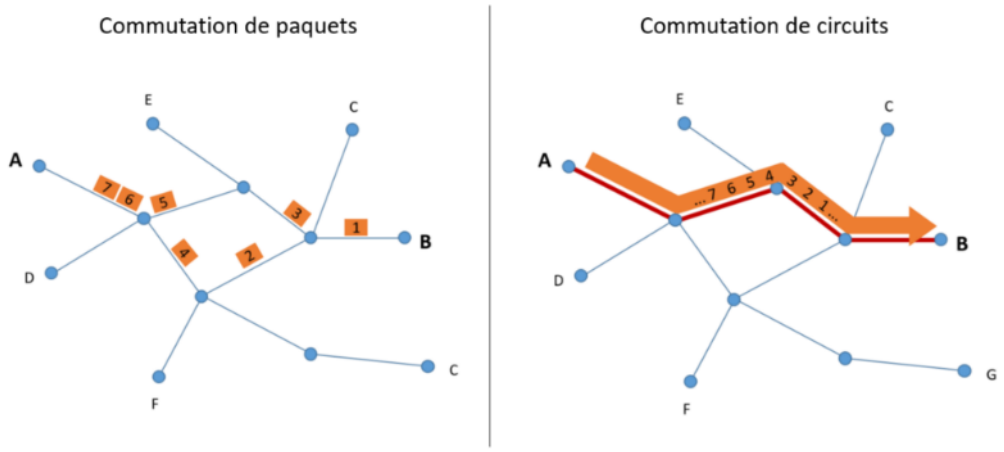


Figure 7 : Transfert d'informations entre deux nœuds A et B

La commutation par paquets est beaucoup plus souple et adaptée à un réseau très sollicité. Cependant, elle présente un inconvénient majeur : sa latence (temps de propagation de l'information) est supérieure à celle d'un réseau par commutation, puisqu'il faut gérer le système de paquets, en particulier la potentielle perte de paquets, qui peut se produire dans le cas où un nœud du réseau est surchargé. Pire, cette latence n'est pas constante, puisqu'elle évolue en permanence en fonction des routes disponibles pour acheminer les paquets. Cette variation de la latence est appelée gigue, et est très problématique pour les applications temps-réel, qui cherchent toujours à minimiser la latence.

Le réseau internet n'étant pas conçu initialement pour la télécommunication, des protocoles et applications spécifiques ont dû être inventés. Les applications de VoIP (voix sur IP), permettent de passer des appels téléphoniques : une des plus connue est Skype, mais il en existe bien d'autres (Messenger, Discord, Viber, WhatsApp…). Il existe également une interface de programmation gratuite dédiée à la vidéoconférence : WebRTC (Web Real-Time Communication). C'est cette interface qui est utilisée pour notre robot de téléprésence, Robair. Nous donnerons donc dans la suite quelques chiffres associés aux performances de WebRTC. Les deux grandeurs qui nous intéressent sont la quantité d'informations pouvant être transmises à chaque instant, et la latence, ou temps de propagation de l'information.

1.2.1.2 Quantité d'informations limitée

La quantité d'informations échangées à chaque instant est limitée par le débit binaire local. Or ce débit binaire dépend du réseau utilisé (cf. Tableau 1). De plus, il doit être partagé entre les différents utilisateurs du réseau. Ainsi, si le débit maximal de la Wi-Fi 4G est théoriquement de 150 Mbit/s, l'entreprise NPerf, spécialisée dans l'étude des performances des réseaux, a mesuré en 2019 un débit binaire moyen de seulement 10 Mbit/s (NPerf, 2020).

Tableau 1 : Ordre de grandeurs pour les débits binaires en fonction du type de réseau utilisé (source : Wikipédia)

Type de réseau	Nom du réseau	Débit binaire théorique
Réseau filaire	ADSL	1 Mbit/s
	FFTH (fibre optique)	300 Mbit/s
Réseau sans fil	Bluetooth	2 Mbit/s
	Wi-Fi 2G	10 kbit/s
	Wi-Fi 3G	2 Mbit/s
	Wi-Fi 4G	150 Mbit/s
	Wi-Fi 5G	10 Gbit/s

*Remarque : Il existe en réalité deux débits différents : le **débit montant**, correspondant aux informations transmises vers le réseau, et le **débit descendant**, correspondant aux informations transmises vers l'utilisateur. Lorsqu'un particulier s'abonne à un réseau, le débit descendant est généralement plus élevé que le débit montant. Les débits indiqués dans le tableau correspondent donc au débit le plus faible.*

Si le débit binaire est trop faible, le seul moyen de conserver la communication en temps-réel est de réduire la quantité d'information transmise, en dégradant la qualité des flux audio et vidéo. Dans un cas critique, le robot peut se trouver dans une zone sans-Wifi et le pilote ne peut plus ni voir ou entendre ses interlocuteurs, ni contrôler le robot. À titre indicatif, pour une application basée sur WebRTC, (Jansen et al. 2018) ont constaté qu'un débit binaire d'au moins

250 kbit/s était nécessaire pour pouvoir maintenir un flux vidéo acceptable à la plus faible résolution possible (480x270 px). Si les réseaux Wifi de cinquième génération (5G) promettent des débits très supérieurs aux débits actuels, la 5G est très décriée pour son impact environnemental, et ses enjeux géopolitiques (Schneidermann, Gramaglia 2020).

1.2.1.3 Temps de propagation non négligeable

Par ailleurs, la latence, ou temps de propagation, a un effet négatif sur les télécommunications. Cet effet a été modélisé par l'Union Internationale des Télécommunications (ANON. 2015). En particulier, l'UIT propose un schéma simplifié qui permet de prédire l'impact de la latence sur la qualité de la communication (cf. Figure 8). Les temps de propagation « bouche à oreille » indiqués correspondent au temps nécessaire pour que la parole d'un utilisateur atteigne l'oreille de son interlocuteur. Leur étude préconise un temps de propagation inférieur à 100 ms, tout en soulignant que certaines applications hautement interactives risquent tout de même d'être affectées par de tels retards. Elle fixe une limite maximale de temps de propagation acceptable à 400 ms.

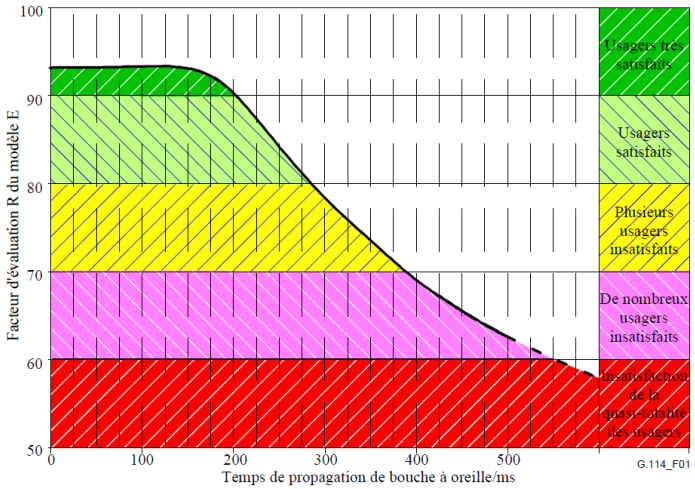


Figure 8 : Détermination des effets du temps de propagation absolu selon le modèle E (ANON. 2003a)

Concernant les applications basées sur WebRTC, (Počta, Komperda 2016) ont mesuré une latence moyenne de 155 ms pour un appel en vidéoconférence. (Jansen et al. 2018) fournissent des mesures plus optimistes : 78 ms de délais dans le cas d'une communication intercontinentale entre New York (côté est des Etats-Unis) et l'Oregon (côté ouest des Etats-Unis), et seulement 215 ms de délai pour un appel extracontinental, entre New York et Sydney (Australie). Dans les deux cas, il s'agit de mesures effectuées entre des appareils connectés à un réseau filaire. Dans ces conditions, WebRTC parvient donc à maintenir une latence acceptable pour de la visioconférence. Cependant, en cas de connexion par réseau Wifi, le délai n'est pas constant, ce qui occasionne des chutes dans le nombre d'images par secondes affichées qui nuisent à la qualité vidéo (Jansen et al. 2018).

1.2.2 Conséquences sur le « toucher social »

La quantité d'informations échangées entre le pilote et le robot de téléprésence est nécessairement limitée. Le pilote n'a donc pas accès à un enregistrement haute-définition de l'environnement du robot. De plus, à cause de la latence, il est condamné à percevoir et agir avec un temps de retard. Ces deux contraintes ont un impact direct sur le « toucher social » du téléopérateur : c'est-à-dire, sa capacité à contrôler finement ce qu'il veut savoir, faire et montrer à ses interlocuteurs. Dans cette partie, nous étudierons plusieurs problèmes récurrents en robotique de téléprésence, et présenterons les solutions proposées dans la littérature.

1.2.2.1 Effet d'écho pour le locuteur

Non seulement la latence nuit à la télécommunication, mais elle engendre également un effet d'écho qui peut être extrêmement gênant pour les utilisateurs. En effet, comme le haut-parleur et le microphone utilisés sont généralement proches l'un de l'autre, la voix qui sort du haut-parleur est enregistrée une seconde fois par le microphone : l'utilisateur peut donc entendre un écho de sa propre voix, décalé dans le temps.

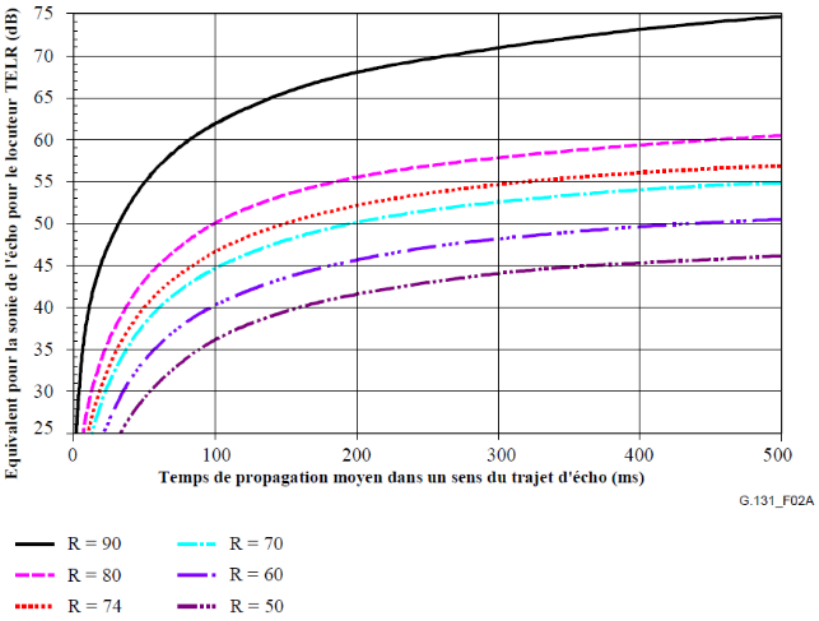


Figure 9 : Effets de l'écho pour le locuteur déterminés à partir du modèle E
(ANON. 2003b)

Cet écho fait l'objet d'une norme UIT, qui estime que l'impact sur la télécommunication dépend à la fois du temps de propagation et de la différence de niveau entre la voix originale et son écho, caractérisée par le TELR (*talker echo loudness rating*). Ainsi, plus le temps de propagation est grand, plus le TELR doit être élevé pour pouvoir maintenir la qualité de communication. Le modèle proposé par l'UIT est représenté Figure 9 : chaque courbe correspond à un facteur de qualité R. C'est le même facteur R qui apparaissait en ordonnée sur la Figure 8, dans le modèle prédisant l'impact du temps de propagation. Par exemple, pour avoir une communication de très bonne qualité avec un temps de propagation de 100 ms, le TELR doit être d'au moins 65 dB. Autrement dit, l'écho doit être imperceptible.

Pour augmenter drastiquement le TELR, la solution la plus simple consiste à équiper le pilote du robot de téléprésence d'un casque audio. Ainsi, on s'assure que la voix de ses interlocuteurs ne fera pas écho dans l'environnement distant, car la voix qui sort du casque audio est trop faible pour pouvoir être enregistrée par le microphone du pilote. En revanche, il paraît difficile d'équiper tous les interlocuteurs potentiels de casques ou d'écouteurs. Il faut donc se tourner vers des solutions algorithmiques. La plus simple consiste à utiliser un canal de communication *half-duplex* : pendant que l'un parle, le microphone de l'autre est systématiquement coupé, ou du moins, son intensité est fortement réduite. L'inconvénient de cette méthode est qu'elle empêche tous chevauchements de parole, qui sont pourtant courants lorsque plusieurs personnes communiquent en chair et en os.

D'autres solutions plus complexes peuvent être mises en œuvre : elles consistent à filtrer le signal reçu, connaissant le signal envoyé, afin de réduire au maximum l'écho. Autrement dit, il faut parvenir à estimer le "chemin" parcouru par l'écho, pour pouvoir le retrancher. Or ce "chemin" est non seulement difficile à modéliser, mais également variable au cours du temps, puisque le robot peut se déplacer, donc les propriétés acoustiques de son environnement sont susceptibles d'évoluer. La question du temps-réel se pose également, puisque le filtrage à appliquer doit être calculé en un temps limité à partir de données limitées. En conséquence, il n'existe pas d'annulateur d'écho acoustique parfait, qui permette d'extraire à partir du signal reçu un signal dénué d'écho. En pratique, il faut donc trouver un compromis entre l'intensité du signal émis, la part d'écho annulée, et la qualité du signal après annulation d'écho (y compris lorsqu'il y a chevauchement de parole) (Enzner et al. 2014).

1.2.2.2 Effet « faux jeton »

Réussir à transmettre l'information du regard est une autre grande problématique de la robotique de téléprésence. En effet, le regard est un élément essentiel de l'interaction, et fondamentalement différent de l'ouïe ou du toucher, comme l'explique Charles Lenay[8] dans une interview réalisée dans le cadre du labex Arts-H2H[9] (Lenay 2016). Le chercheur, qui s'intéresse aux croisements perceptifs, expliquent que les sensations acoustiques et tactiles sont le plus souvent partagées entre les personnes présentes dans un même lieu : si je peux percevoir quelqu'un en le touchant, ou en l'entendant, alors lui aussi me touche et peux m'entendre. Au contraire, la vision permet de percevoir quelqu'un sans qu'il nous perçoive, et de savoir que telle personne voit ce que telle autre ne voit pas. Pour illustrer son propos, Lenay prend un exemple du quotidien : une fois qu'on a croisé le regard de quelqu'un qu'on connaît, on ne peut

[8] Chercheur à l'Université de Technologie de Compiègne, il s'intéresse aux croisements perceptifs, en particulier tactiles (Lenay 2010)
[9] Laboratoire d'excellence des arts et médiations humaines

plus faire croire qu'on ne l'a pas vu. Il est donc essentiel de parvenir à reproduire ce « toucher visuel » en robotique de téléprésence.

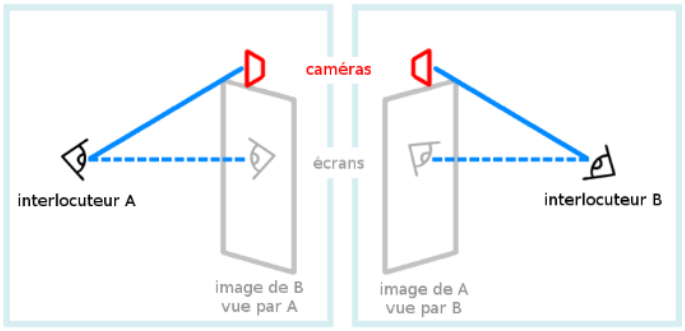


Figure 10 : Illustration du problème du croisement de regards (effet « faux-jeton ») en visioconférence

Hors, en visiophonie classique, le croisement de regard est impossible. En effet, si la caméra est placée en haut de l'écran, l'utilisateur aura toujours l'air de regarder vers le bas, à moins qu'il se force à faire un regard caméra (cf. Figure 10). Ainsi, deux interlocuteurs ne peuvent en aucun cas se regarder dans les yeux simultanément. C'est ce qu'on appelle l'effet « faux-jeton », car en Occident, refuser de croiser le regard d'une personne peut être interprété de façon très négative. Ce problème est peut-être contourné par le robot QB (cf. Figure 11) : le regard de l'interlocuteur étant plus attiré par les « yeux » des caméras que par la vidéo du pilote ; tandis que la vidéo du pilote est trop petite pour pouvoir distinguer la direction de son regard.

Figure 11 : Photos du robot QB
(http://www.roboticamente.net/qb-il-robot-avatar-per-la-presenza-remota-virtuale/)

Il existe également des solutions optiques permettant de corriger ce défaut. Elles ont déjà été testées avec succès pour la visioconférence (Buxton 1992 ; Vertegaal et al. 2003). Elles sont basées sur le principe du prompteur (cf. Figure 12) : entre la caméra et l'utilisateur est placée une vitre semi-réfléchissante ; sous la vitre est placé l'écran où apparaît l'image de l'interlocuteur ; grâce à un jeu d'ombre, la caméra peut filmer tandis que l'utilisateur peut voir l'écran. L'avantage de ces systèmes est que la position optimale pour voir est également la position optimale pour être vu (Buxton 1992) : ainsi les utilisateurs sont intuitivement conscients de ce qu'ils montrent à l'autre. Cependant, ce dispositif est peu adapté à la téléprésence, car il ne fonctionne parfaitement qu'à condition que les utilisateurs se tiennent à une position précise et qui tient compte des sources de lumière ; dans le cas contraire, apparaissent des reflets sur la vitre semi-réfléchissante, ou le contraste n'est pas suffisant pour permettre aux interlocuteurs de se voir. Ainsi le « Floating Avatar » proposé par (Tobita et al.

2011), un petit dirigeable à l'enveloppe semi-transparente embarquant une caméra et un vidéoprojecteur, ne semble fonctionner que dans la pénombre.

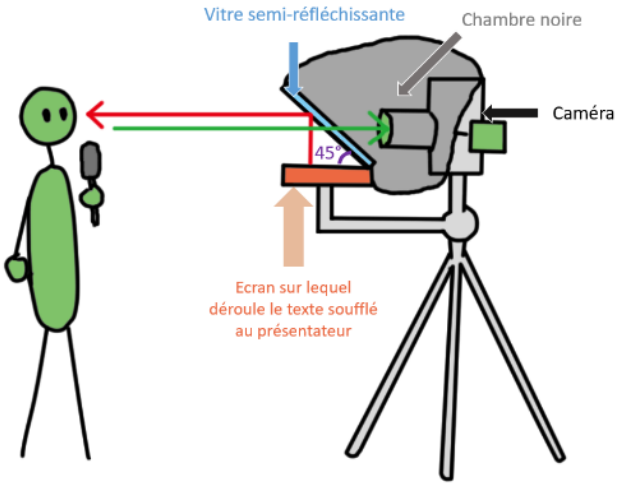


Figure 12 : Schéma de principe du prompteur

Plus récemment, Apple a incorporé un système de correction du regard à son application de vidéoconférence FaceTime (Le Monde, 4 juillet 2019). Il repose sur des algorithmes de modélisation du visage basés sur des réseaux de neurones, et permet de donner l'impression que l'utilisateur regarde droit devant lui lorsqu'il regarde son écran, et vers le haut lorsqu'il regarde sa caméra. Ce système a été vraisemblablement inspiré par la startup allemande CatchEye, qui proposait jusqu'en 2017 un système similaire basé sur la kinect.

1.2.2.3 Navigation difficile

Un des problèmes majeur en robotique de téléprésence est la navigation du robot. En effet, il n'est pas aussi simple de piloter un robot à distance qu'une voiture télécommandée que l'on aurait en ligne de vue directe. Tout d'abord, le point de vue du pilote est un point de vue à la première personne, limité par l'angle de la caméra du robot. En outre, il s'agit le plus souvent d'une image en deux dimensions, donc qui ne permet pas d'apprécier correctement les distances. Troisièmement, du fait du temps de propagation, le pilote agit toujours avec un temps de retard : il est donc incapable de s'adapter à un environnement qui évolue rapidement. Ainsi, il est très difficile en téléprésence de repérer les obstacles et de les éviter. Pour que le robot ne rentre pas dans un mur ou une personne, des systèmes de détection à courte portée sont en général utilisés, pour immobiliser le robot s'il arrive trop près d'un obstacle. Cependant, ces systèmes peuvent bloquer entièrement les mouvements du robot, en particulier si le pilote essaye de passer par une ouverture étroite.

Faciliter le pilotage du robot est donc d'abord un enjeu de sécurité. Cela peut être également une source de frustration et une charge mentale pour le pilote, lorsque celui-ci n'a pas de difficulté pour se déplacer en temps normal, et doit réduire sa vitesse au minimum pour pouvoir faire naviguer le robot d'une pièce à une autre. S'il éprouve de grandes difficultés à manipuler le robot, il ne pourra pas se concentrer sur ses interactions sociales. Il serait donc bénéfique

d'automatiser les mouvements du robot, afin qu'il puisse réaliser les mouvements attendus par le pilote sans que celui-ci ait besoin d'entrer les commandes exactes. De tels automatismes sont déjà implémentés sur certains robots de téléprésence. Ainsi, les robots Beams intègrent un système d'aide au parking automatique : le pilote n'a qu'à appuyer sur un bouton pour que le robot aille se placer sur son socle de chargement. Pour les dernières versions du robot Double, un système de navigation automatique permet à l'utilisateur de cliquer sur l'emplacement au sol où il souhaite que le robot se dirige, et celui-ci va s'y placer en évitant les obstacles éventuels (cf. vidéos de démonstrations du constructeur).

Cependant, la navigation d'un robot ne peut pas se limiter à de l'évitement d'obstacles : en général, le robot évolue dans un milieu vivant, traversé par des personnes qui ne sont pas immobiles, et risquent de réagir aux mouvements du robot. Il y a donc des règles sociales à respecter : ne pas couper la route, se tenir à une certaine distance etc. Or, a priori, ces règles ne sont pas forcément les mêmes pour un robot téléopéré que pour un humain en chair et en os, car ils n'ont pas le même corps et ne sont pas perçus de la même façon.

1.2.2.4 Portée vocale dégradée

Une autre spécificité de la téléprésence est qu'il est possible de régler le volume sonore. En particulier, en modifiant le volume du robot, il est possible de modifier artificiellement sa portée vocale. Malheureusement, cette portée vocale n'est pas toujours adaptée à la situation, ce qui peut entraîner des erreurs communicationnelles : par exemple, la voix qui sort du robot est soudain si forte qu'elle surprend ou dérange les interlocuteurs. Au contraire, si le volume est trop faible, le pilote n'arrivera pas à capter l'attention de ses interlocuteurs, qui risquent de l'ignorer. Il faut donc trouver des solutions pour régler ce volume de façon satisfaisante.

Figure 13 : Illustration de l'impact du volume sonore du robot sur l'interaction.

La solution la plus simple consiste à laisser le réglage du robot aux personnes présentes dans l'environnement local. Cependant, cette solution n'est pas forcément acceptable par le pilote, qui est alors dépendant de ses interlocuteurs. En outre, comme le robot de téléprésence sert d'avatar à son pilote, il peut paraître indélicat de le manipuler comme un simple écran de télévision. Ainsi, (Paepcke et al. 2011) ont constaté que même lorsque les interlocuteurs ont la possibilité de modifier le volume du robot, ils ne le font pas en général. Ils auraient également des scrupules à demander au pilote de baisser le volume de son microphone, car il semble impoli de demander à quelqu'un de parler moins fort.

D'autres solutions ont donc été proposées dans la littérature pour répondre à ce problème. Celle de (Paepcke et al. 2011) s'appuie sur une astuce utilisée initialement en téléphonie : les ***sidetones***. Il s'agit de réintroduire artificiellement dans le signal entendu par l'utilisateur un retour de sa propre voix. Contrairement à l'écho provoqué par la téléprésence ($ 1.2.2.1), ce retour est quasi instantané, et ne dérange donc pas le pilote. Cependant, son effet est assez limité : si le retour est trop faible (- 3 dB), on n'observe pas de différence par rapport à la condition sans retour ; s'il est assez fort (+ 1.7 dB), le pilote du robot a tendance à parler légèrement moins fort (d'environ - 1 dB). Ces résultats ont été obtenus pour une écoute au casque : dans le cas de l'utilisation de haut-parleurs, (Paepcke et al. 2011) n'ont pas pu observer d'effet significatif des *sidetones*. Ainsi, l'utilisation d'un retour acoustique peut inciter le pilote à parler moins fort, mais ne permet donc en aucun cas de régler le volume sonore du robot dans le cas où celui-ci ne serait pas approprié.

Une autre méthode utilisée en robotique de téléprésence est celle du **contrôle automatique de gain**. Elle consiste à modifier le volume du robot, de sorte que la voix qui en sort soit à la bonne intensité. En première approximation, on peut ainsi décider arbitrairement que l'intensité en sortie du robot doit être de 60 dB(A). Ensuite, pour pouvoir s'adapter au bruit environnant, on peut imaginer fixer le rapport signal-sur-bruit. Dans un troisième temps, il est possible de tenir compte de la distance entre le robot et l'interlocuteur pour tenir compte du fait que le rapport signal-sur-bruit diminue avec la distance. C'est un système de ce type que proposent (Hayamizu et al. 2014), en s'appuyant sur des courbes de pondération obtenues lors de tests utilisateurs (cf. Figure 14). Le robot, équipé d'une kinect, détecte le visage des interlocuteurs. Le pilote peut alors sélectionner la personne à qui il parle, et le volume de sa voix est fixé en fonction du bruit ambiant et de la distance mesurée. Le pilote a le choix entre deux réglages : un mode « confortable » qui permet à son interlocuteur de l'entendre sans tendre l'oreille, et un mode « secret », qui fait que seule la personne la plus proche du robot doit pouvoir entendre ce qui est dit.

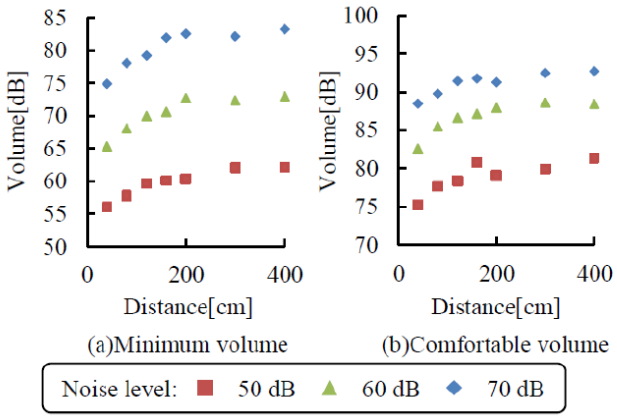


Figure 14 : Volume optimal d'un robot de téléprésence en fonction du bruit ambiant et de la distance de l'interlocuteur (Hayamizu et al. 2014).

Une troisième catégorie de méthodes consiste à essayer de donner au pilote les informations nécessaires pour qu'il contrôle lui-même le volume du robot. Cela peut passer par une interface graphique, permettant au pilote de visualiser sa portée vocale. De telles interfaces ont déjà été

développées pour la visioconférence par (Kimura et al. 2007), qui s'intéressent à des systèmes de télécommunication actifs en permanence (Kimura et al. 2006). Leur objectif est de permettre à une personne d'initier poliment une conversation à distance de façon moins intrusive qu'un coup de téléphone, en contrôlant d'abord si la personne est disponible grâce à la vidéo, puis en essayant d'attirer son attention en parlant à une intensité appropriée au contexte. Ils proposent d'équiper l'interlocuteur d'un microphone calibré, afin de pouvoir mesurer l'intensité acoustique qui parvient à son niveau, et l'indiquer ensuite au pilote. Leur premier prototype consiste à afficher cette intensité sur l'écran du pilote, sous formes de barres qui rayonnent depuis le haut-parleur (cf. Figure 15). Plus le nombre de barre est élevé, plus la voix est forte. Si au moins une barre apparaît au niveau de l'interlocuteur, alors le volume est assez fort. Afin de rendre cet affichage plus clair pour le pilote, et visible par l'interlocuteur, le second prototype consiste à afficher ces barres dans l'environnement local en les projetant au sol à l'aide d'un vidéo-projecteur. Le troisième prototype consiste en un disque de leds, reliées à un microphone : elles s'éclairent lorsque l'intensité acoustique mesurée dépasse un certain seuil. Ainsi, lorsque l'utilisateur voit les leds s'éclairer du côté de son interlocuteur, il est certain de parler suffisamment fort pour être entendu.

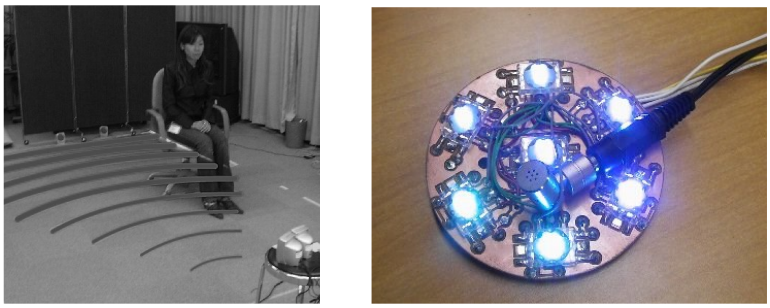


Figure 15 : Deux systèmes de visualisation de la portée vocale proposés par (Kimura et al. 2007)

1.2.2.5 Effet superman

La dernière problématique que nous allons évoquer concerne la manière dont le téléopérateur distant est perçu par ses interlocuteurs à travers le robot. (Stoll et al. 2018) se sont intéressés à l'usage de robots de téléprésence pour le travail collaboratif. Ils ont étudié des groupes de trois personnes devant résoudre une tâche de décodage de mots : deux étaient présentes dans l'environnement local, et une téléprésente. Trois conditions étaient étudiées. En condition 1, seul le pilote du robot disposait des informations de décodage. En condition 2, tous les participants disposaient de l'information. En condition 3, seuls les sujets présents dans la pièce disposaient de l'information. Les chercheurs ont constaté que les personnes téléprésentes participaient moins aux interactions. En particulier, dans les cas où les sujets présents possèdent déjà l'information nécessaire pour résoudre la tâche, ils risquent d'exclure la personne à distance en travaillant en binôme. Les chercheurs ont également constaté que les participants locaux faisaient moins confiance à leur collaborateur distant, qu'à celui présent dans la pièce. Dans leur article, ils soulignent l'ambiguïté de ce dernier résultat : Est-ce que les sujets font moins confiance au robot, dont ils ignorent le fonctionnement, ou à son pilote ?

Ainsi, l'image du pilote du robot aux yeux de ses interlocuteurs pourrait être déformée par la téléprésence, en particulier s'ils ne se sont jamais rencontrés en chair et en os. Comme il participe moins aux discussions, il pourrait perçu comme plus timide, ou moins compétent qu'il ne l'est habituellement. Le décalage serait encore plus important si le pilote du robot choisissait de modifier son apparence visuelle ou vocale, ou si le robot augmentait ses capacités physiques. Dans un cas extrême, on pourrait imaginer qu'une personne n'ait pas les mêmes relations sociales lorsqu'elle pilote un robot de téléprésence, que lorsqu'elle est présente en chair et en os. Ce phénomène encore hypothétique a été baptisé « effet Superman » par Véronique Aubergé, en référence au personnage de Lois Lane. La jeune femme, quoique très intelligente, tarde à accepter le fait que Superman et Clark Kent sont une seule et même personne : à ses yeux, le premier est un superhéros, qu'elle trouve particulièrement attirant, tandis que le deuxième est un simple collègue de bureau, maladroit et ordinaire.

1.2.2.6 Résumé

La robotique de téléprésence est un nouveau champ des télécommunications, qui ambitionne de pouvoir téléporter une personne d'un lieu à l'autre. Pensée initialement pour l'exécution de tâches techniques à distance en environnement hostile, la téléprésence apparaît aujourd'hui d'abord comme un moyen d'interagir avec d'autres personnes, notamment dans des situations où le lien social est dégradé. L'enjeu n'est donc plus de simuler uniquement la présence physique du téléopérateur, mais également sa présence sociale.

Or, ces deux types de présences ne sont pas suffisamment transmises par les robots de téléprésence actuels, y compris dans le cas des prototypes les plus perfectionnés. Le pilote du robot ne perçoit pas l'environnement local comme s'il y était : il n'a accès qu'au son et à l'image, dans une résolution faible. À cause de la latence, cette perception est le plus souvent figée : elle n'évolue pas de façon instantanée en fonction des mouvements de tête du pilote. D'ailleurs, le robot de téléprésence ne peut pas reproduire exactement les gestes du pilote, car il n'a pas la motricité nécessaire. Ce manque de présence physique a nécessairement des conséquences sur la présence sociale : les interlocuteurs n'ont pas l'impression de s'adresser à une vraie personne, mais à un robot téléopéré. Dans le pire des cas, ces limites techniques peuvent engendrer des artefacts socio-affectifs, aux conséquences bien réelles : peu importe que le pilote du robot soit responsable ou non des erreurs commises en téléprésence (ex : couper la parole, ne pas céder le passage, parler trop fort…), elles vont influencer sa relation aux autres.

1.3 Une marge de progression pour le « toucher vocal »

Dans cette thèse, nous nous focaliserons uniquement sur une partie des signaux sociaux qui participent au « toucher social » : les signaux vocaux. Autrement dit, il s'agit de permettre au pilote du robot de téléprésence de contrôler ce qu'il peut entendre, et faire entendre à ses interlocuteurs. Ce toucher vocal est fondamental en communication parlée : ainsi, il peut exister de la visioconférence sans image, mais pas sans son. Or, nous verrons dans cette partie que la question du toucher vocal a l'air de peu intéresser les entreprises : elles se concentrent plutôt sur les capacités visuelles et motrices des robots.

1.3.1 Peu d'informations sur le matériel et les fonctionnalités audio

La Figure 16 récapitule les informations obtenues à la lecture des notices des 28 robots de téléprésence présentés dans le catalogue de la section 1.1.2. Nous nous sommes intéressés à certaines fonctionnalités de ces robots, qui concernent directement le toucher vocal, visuel, et la navigation. Les barres vertes correspondent aux informations renseignées dans les notices. Les barres bleues correspondent aux informations non pertinentes, soit parce qu'elles ne concernent pas ce type de robot (ex : un robot à base fixe n'a pas besoin de détecteur d'obstacles), soit parce que le robot est fourni sans écran, et que ces informations dépendent donc de la tablette/smartphone utilisé. Enfin, les barres jaunes correspondent aux informations qui ne sont pas renseignées dans la notice. Dans le cas de la détection d'obstacles, ou du parking automatique (robot qui se branche tout seul à la borne de recharge électrique), on peut supposer que lorsque l'information n'est pas renseignée, c'est que l'option n'est pas disponible sur le robot. En revanche, on sait qu'un robot de téléprésence sera toujours équipé d'au moins une caméra, un écran, un microphone et un haut-parleur ; dans ce cas, il s'agit d'un manque d'informations.

On constate que les caractéristiques qui concernent le toucher visuel et la navigation sont mieux renseignés en moyenne que celles qui concernent le toucher vocal. Cela se vérifie à la fois pour le matériel et les fonctionnalités. Ainsi, en mettant de côté les décomptes des informations non pertinentes, 70% des notices contiennent des informations concernant la ou les caméras utilisées. Les descriptifs techniques disponibles en ligne insistent sur la qualité des caméras utilisées : elles sont haute définition, grand angle, et peuvent parfois être pilotées à distance, pour zoomer, ou changer leur angle d'inclinaison. Au contraire, les informations concernant les microphones et les haut-parleurs ne sont renseignés respectivement que dans 55% et 39% des cas. Par ailleurs, en éliminant les robots de téléprésence à base fixe, la majorité des modèles intègrent des fonctionnalités de navigation : 71% font de la détection d'obstacles, et 50% peuvent se parquer automatiquement. Au contraire, seuls 25 % des notices mentionnent une annulation de bruit, et 17% évoquent la question du contrôle du volume (automatisé, ou accessible facilement aux interlocuteurs).

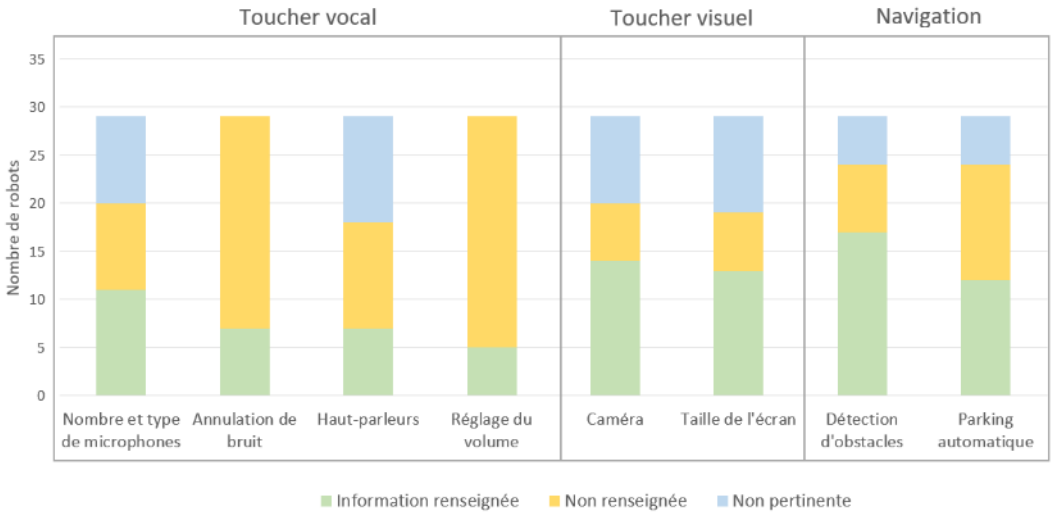


Figure 16 : Analyse des notices de robots de téléprésence (cf. Tableau Annexe A)

1.3.2 L'intelligibilité, plutôt que l'immersion

Voici quelques exemples de descriptifs lisibles sur les sites officiels d'entreprise en robotique de téléprésence :

> Awabot : « Les performances audio et vidéo du dispositif de téléprésence mobile BeamPro sont adaptées aux environnements hostiles, qu'ils soient vastes, fortement fréquentés ou bruyants. »
>
> Double Robotics (traduit de l'anglais): « Un champ sophistiqué de six microphones permet au pilote d'entendre les gens de plus loin, et avec moins de bruit de fond. Le système audio intégré permet à la communication audio full-duplex (audio simultannée en deux voies) d'être plus robuste en environnement difficile. »
>
> OhmniLabs (traduit de l'anglais) : « Microphone pour champ lointain et haut-parleur. Entendre et être entendu. Un haut-parleur Jabra intégré pour un maximum de performances audio. »
>
> VGo (traduit de l'anglais) : « Quatre microphones pour saisir ce que vous avez besoin d'entendre. Un haut-parleur haut pour un son précis. Un haut-parleur bas pour un son complet. »

Ainsi, lorsque l'équipement audio est mis en avant, c'est pour insister sur la possibilité de communiquer même en environnement bruyant. Les microphones utilisés permettent de faire de l'annulation de bruit, ou de capter le son à distance ; tandis que les haut-parleurs doivent être assez puissants pour porter la voix du pilote. Le problème traité est donc celui de l'intelligibilité, qui doit être maintenue quelles que soient les conditions dans l'environnement local. La question de l'immersion acoustique, elle, n'est pas évoquée.

1.3.3 Résumé

On trouve relativement peu de renseignements concernant le matériel audio utilisé, ou les fonctionnalités audio des robots de téléprésence ; comme si ces informations étaient secondaires. Dans les rares cas où elles sont présentes, ce n'est pas l'immersion acoustique qui est mise en avant, mais l'intelligibilité : il ne s'agit pas de permettre au pilote d'entendre l'environnement comme s'il y était, mais d'assurer que sa voix, et celles de ses interlocuteurs, se détachent du fond sonore. Pourtant, nous verrons dans la partie suivante qu'il existe aujourd'hui des solutions pour permettre une meilleure immersion acoustique à distance.

1.4 Des technologies pour l'immersion acoustique en téléprésence

Dans cette dernière partie, nous nous intéressons aux technologies permettant de réaliser une immersion acoustique à distance. Comme annoncé en section 1.1, cette immersion est particulièrement importante pour susciter le sentiment de téléprésence ; ainsi que pour permettre au pilote du robot de contrôler son toucher vocal.

Nous commencerons par expliquer comment réaliser des oreilles artificielles, qui permettent à un auditeur d'entendre un environnement distant à l'aide d'un simple casque audio. Nous évoquerons ensuite les bouches artificielles, qui permettent de reproduire l'acoustique d'une voix humaine. Enfin, nous verrons quelques exemples d'application de ces technologies en robotique de téléprésence.

1.4.1 Oreilles artificielles

L'idée d'oreilles artificielles est très simple et presque aussi vieille que celle du téléphone. Ainsi, à l'exposition universelle de 1881, Clément Ader fait placer une paire de microphones de chaque côté de la scène de l'Opéra Garnier, et une autre à la Comédie française. Ces oreilles artificielles sont reliées par câble téléphonique à des paires d'écouteurs mises à disposition dans le Palais de l'industrie (Gasiglia-Laster 1983).

Figure 17 : Le théâtrophone, affiche de (Chéret 1896)

Si la qualité de la retransmission laisse encore à désirer, la scène acoustique apparaît en relief : grâce à ces deux canaux audio, il est possible de percevoir les déplacements des acteurs sur scène, ou de localiser un rire dans le public. Ce « théâtrophone » connaît un franc succès, comme en témoigne une note de Victor Hugo :

> *C'est très curieux. On se met aux oreilles deux couvre-oreilles qui correspondent avec le mur, et l'on entend la représentation de l'Opéra, on change de couvre-oreilles et l'on entend le Théâtre-Français, Coquelin, etc. On change encore et l'on entend l'Opéra-Comique. Les enfants étaient charmés et moi aussi.*
>
> *(Hugo, 1970)*

Aujourd'hui, cette manière de créer une immersion sonore à partir de deux signaux acoustiques est qualifiée de binaurale (« deux oreilles »). Un son binaural peut être obtenu soit par l'enregistrement, soit par la simulation ; mais avant d'aborder ces aspects, il est nécessaire de comprendre comment un auditeur perçoit acoustiquement l'espace. En effet, le principal intérêt du son binaural, et ce qui le rend immersif, c'est qu'il permet à l'auditeur de localiser dans l'espace différentes sources sonores.

1.4.1.1 Localisation spatiale des sources sonores

Il existe une littérature abondante en psychoacoustique concernant la localisation sonore, c'est-à-dire la capacité des auditeurs à repérer la position des sources sonores dans l'espace. Cette position est décrite dans un référentiel centré sur l'auditeur dont les trois dimensions sont : l'**azimut** (angle horizontal), l'**élévation** (angle vertical) et la **distance** (cf. Figure 18). Par

exemple, une source située face à l'auditeur, à la même hauteur que ses oreilles, et à une distance de 2 m serait repérée par les coordonnées (0°, 0°, 2m).

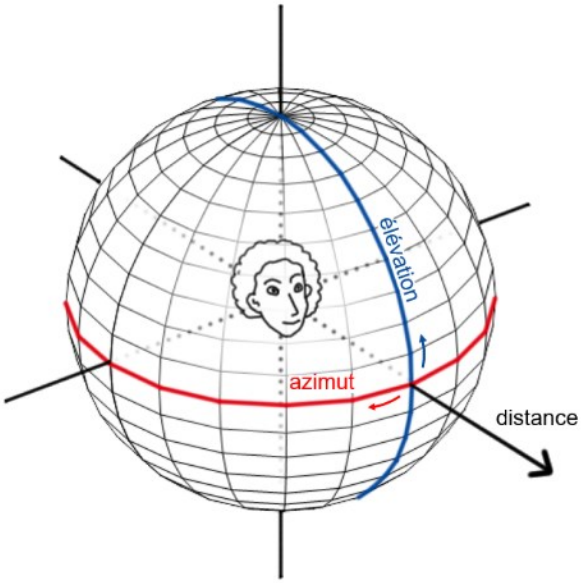


Figure 18 : Repère de localisation des sons

Il existe plusieurs indices acoustiques, qui, associés aux connaissances a priori de l'auditeur, lui permettent de percevoir des informations sur son environnement physique et social. Nous allons voir les principaux, et discuter de leurs contributions respectives.

- Indices acoustiques

Dès le début du XXème siècle, (Lord Rayleigh 1907) propose la Théorie Duplex pour expliquer comment il est possible de percevoir l'azimut d'une source sonore. Il définit deux grandeurs pour distinguer le signal reçu par l'oreille gauche, et celui reçu par l'oreille droite: la **différence de temps inter-oreilles** (ITD : *interaural time difference*) et la **différence d'intensité inter-oreilles** (ILD : *interaural level difference*). À partir de ces deux grandeurs, il est possible de deviner la direction d'arrivée du son. En effet, si le son vient de face, l'ITD et l'ILD sont nuls. En revanche, s'il vient de la gauche, alors il n'arrive pas en même temps aux deux oreilles : l'ITD est positive, car il arrive d'abord à l'oreille gauche, puis à l'oreille droite. Le son perd également en intensité entre l'oreille gauche et l'oreille droite : l'ILD est donc positive.

Ces indices peuvent être qualifiés de **binauraux**, car ils nécessitent deux oreilles. Cependant, ils ne permettent pas d'expliquer la manière dont l'élévation est perçue. Dans les années suivantes, une troisième source d'indices acoustiques a donc été mise en évidence : les **indices monoraux**. Il s'agit de modifications spectrales du son liées à la forme du pavillon, de la tête et du torse (Roffler, Butler 1968).

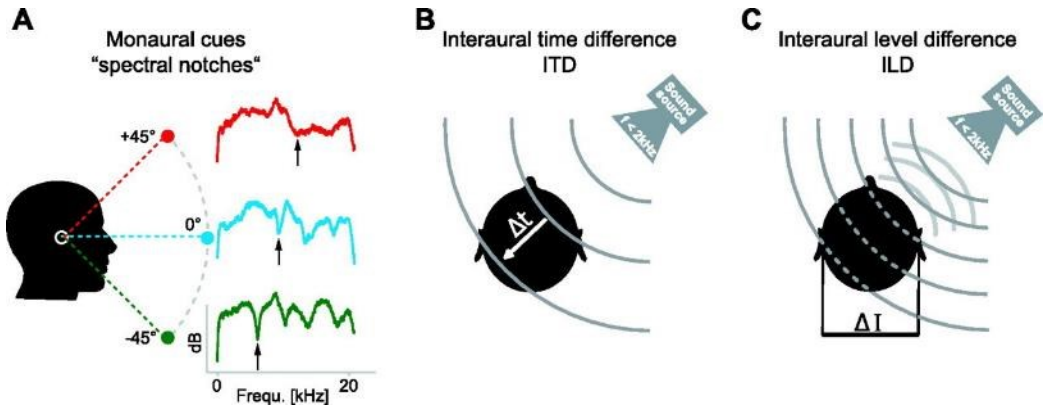


Figure 19 : Principaux indices acoustique permettant la localisation spatiale (Grothe et al. 2010)

Les recherches sur la perception de la distance, plus rares, ont permis de mettre en évidence l'existence d'autres indices spécifiques (Zahorik 2005). Le plus évident est l'**intensité**, puisque celle-ci décroît lorsqu'on s'éloigne de la source sonore. En l'absence de réverbération, cette diminution d'intensité est de 6 dB par doublement de la distance. Le **spectre sonore** est également affecté à très grande distance, puisque les hautes fréquences sont plus absorbées que les basses fréquences (de l'ordre de quelques dB par 100 m). Enfin, dès que l'écoute se fait à l'intérieur d'un bâtiment, l'auditeur peut se fier au **ratio d'énergie directe / réverbérée**. En effet, en présence de surfaces réfléchissantes, le son se réverbère, créant de nouvelles sources sonores secondaires, dont l'intensité cumulée finit par dépasser celle du son direct (cf. Figure 20).

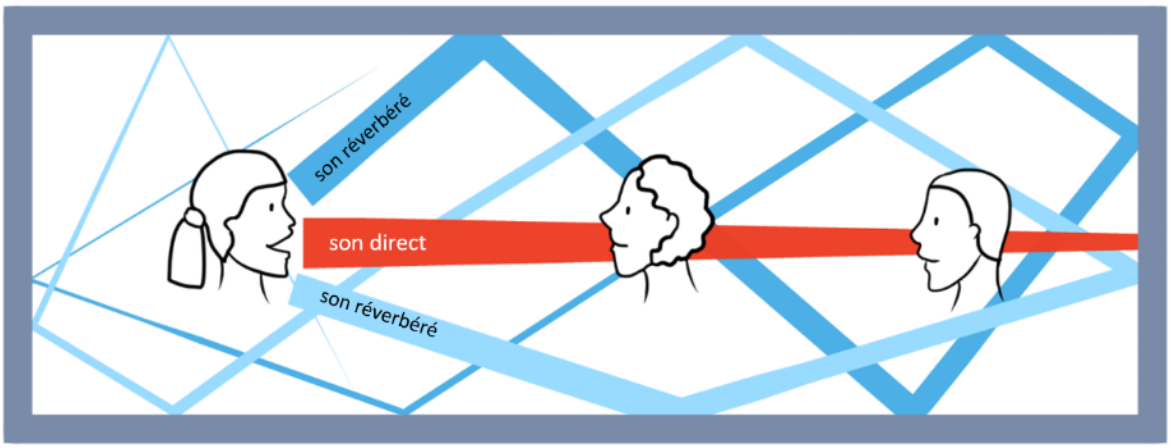


Figure 20 : Illustration du phénomène de réverbération

- Contribution des différents indices

Les différents indices acoustiques que nous venons de présenter contribuent chacun à la perception spatiale des sources sonores. Cependant, leur importance varie. Ainsi, si les indices binauraux, ITD et ILD, permettent de distinguer si le son vient de la gauche ou de la droite, ils ne suffisent pas à déterminer si le son vient de l'avant ou de l'arrière, du haut ou du bas. En effet, ces deux indices sont constants le long d'un **cône de confusion**, et sur le plan médian (cf. Figure 21). Il existe donc des **confusions avant/arrière**, mais qui peuvent être facilement

résolues à partir du moment où l'auditeur peut tourner la tête, ce qui permet de déplacer le cône de confusion.

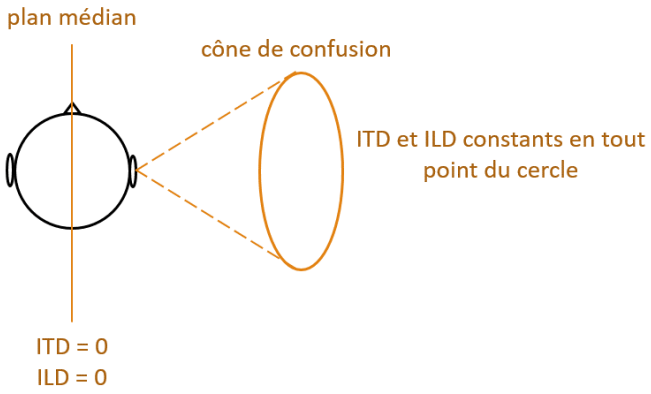


Figure 21 : Sources des confusions avant/arrière

Les performances en localisation dépendent également du type de son utilisé. En effet, les indices acoustiques n'ont pas la même importance, selon que le son est aigu ou grave. Ainsi, la différence de temps inter-oreilles (ITD) est surtout utile pour les sons graves (Middlebrooks, Green 1991 ; Carlile 1996). Pour les sons aigus, le signal évolue beaucoup trop vite pour pouvoir estimer une ITD. Une illustration avec des signaux stéréos artificiels est fournie en Figure 22 : les deux signaux ont exactement la même ITD, et la même enveloppe temporelle ; seule leur fréquence fondamentale varie (100 Hz pour le premier, 2000 Hz pour le second). On constate que pour le premier signal, il n'y a aucune ambiguïté pour estimer l'ITD, puisqu'il s'agit simplement d'une différence de phase entre le signal reçu par l'oreille gauche et celui reçu par l'oreille droite. Pour le second signal en revanche, il est impossible de mesurer l'ITD dans les parties stationnaires (entre 10 et 30 ms), car la période du signal est plus courte que l'ITD (environ 0.5 ms contre 0.6 ms).

Au contraire, pour la différence d'intensité inter-oreilles (ILD), ce sont les sons aigus qui sont favorisés (Middlebrooks, Green 1991 ; Carlile 1996). En effet, cette différence d'intensité n'est pas due à la distance entre les deux oreilles, mais au phénomène de diffraction des ondes sonores : lorsqu'une onde rencontre un obstacle de taille très supérieure à sa longueur d'onde, elle est bloquée, tandis que lorsqu'elle rencontre un obstacle de taille inférieure, elle le contourne. Ainsi, les sons aigus sont fortement atténués par la tête ; tandis que les sons graves ne le sont pas. Par exemple, en supposant que la vitesse du son dans l'air est d'environ de 340 m/s, un son pur dont la fréquence est inférieure à 1 kHz a une longueur d'onde supérieure à 34 cm, il peut donc contourner l' « obstacle » de la tête, qui est de dimension inférieure.

Les hautes fréquences sont également essentielles pour les indices monoraux, car du fait des dimensions de l'oreille, les indices acoustiques spectraux liés à la forme du pavillon apparaissent principalement dans les fréquences supérieures à 8 kHz (Best et al. 2005 ; Roffler, Butler 1968). C'est une donnée importante, car en télécommunications et en étude de la parole, il est très fréquent pour limiter la taille des enregistrements d'utiliser une fréquence

d'échantillonnage de seulement 16 kHz ; ce qui signifie que les fréquences supérieures à 8kHz ne sont pas correctement échantillonnées[10].

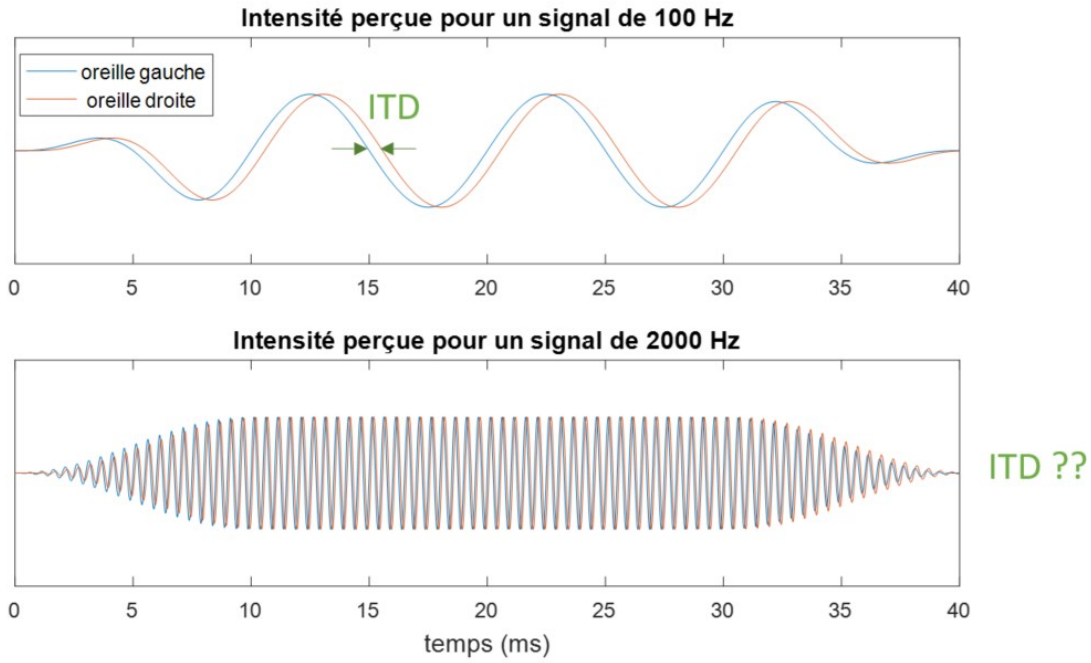


Figure 22 : Illustration de la mesure de la différence de temps inter-oreilles (ITD) pour deux signaux d'intensité différente
Signaux artificiels de même enveloppe temporelle.
ITD correspondant à l'écart entre deux oreilles (environ 20 cm)

Par ailleurs, en ce qui concerne la perception de la distance, nous avons vu qu'il existe deux indices principaux : l'intensité et le rapport d'énergie direct / réverbérée. L'intensité est un indice fiable, mais uniquement à condition d'avoir une intensité de référence à laquelle la comparer : par exemple, lorsque la source se déplace et que son intensité diminue, on devine qu'elle s'éloigne. En revanche, si la distance est fixe, l'auditeur a besoin de faire appel à ses connaissances a priori sur la source sonore, pour deviner à quelle distance correspond l'intensité qu'il perçoit. Si la source sonore ne correspond à rien de connu, il est obligé d'imaginer une intensité de référence.

Au contraire, le rapport d'énergie direct / réverbérée ne nécessite pas d'a priori sur la source sonore. Cependant, il dépend entièrement de l'environnement acoustique, et même de la position de la source sonore et de l'auditeur dans cet environnement. Citons notamment les travaux de (Shinn-Cunningham 2003), qui a étudié l'influence de la position du locuteur à l'intérieur d'une salle de classe virtuelle sur sa perception de l'azimut et de la distance. Quatre positions différentes étaient simulées à partir de mesures acoustiques obtenues à partir d'une tête artificielle KEMAR (cf. Figure 23). L'expérience était divisée en quatre sessions : une pour chaque position simulée. La chercheuse conclut que les performances des sujets sont légèrement meilleures lorsqu'ils se trouvent au centre de la pièce, donc loin des sources de

[10] cf. Théorème de Shannon : « La fréquence d'échantillonnage d'un signal doit être supérieure à deux fois la fréquence maximale du signal. »

réverbération. En outre, ils sont capables de s'adapter à leur environnement, puisque leurs performances s'améliorent au cours des sessions.

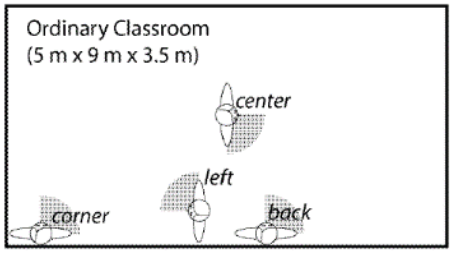


Figure 23 : Schéma des positions de l'auditeur étudiées dans l'article de (Shinn-Cunningham 2003)
Les quadrants représentent la position des sources sonores, issues de combinaisons avec les valeurs suivantes :
azimuts : 0, 45 et 90°
distances : 0.15, 0.40 et 1m

- Autres informations acoustiques

L'ouïe ne permet pas uniquement de repérer des sources sonores dans l'espace ; elle fournit également bien d'autres informations sur l'environnement. Par exemple, le son ne se propage pas de la même manière dans une église, ou dans un champ en extérieur. La présence de surfaces réfléchissantes permet au son de rebondir, ce qui crée des échos. En outre, le son se colore, puisque toutes les fréquences ne se comportent de la même façon : certaines peuvent être amplifiées par l'environnement, ou au contraire fortement atténuées. Il n'existe donc pas de son pur : chaque son, produit dans un milieu donné, porte l'« empreinte acoustique » de ce milieu (Blesser 2007).

En particulier, il est possible d'estimer la taille d'une pièce à partir du temps de réverbération (Yadav et al. 2013). Les échos nous donnent également une indication sur la manière dont elle est meublée : une pièce vide résonne beaucoup plus qu'une pièce encombrée, une cuisine carrelée qu'un salon moquetté. Il est même possible de deviner la présence d'un obstacle, ou d'une ouverture, et d'estimer sa taille (Gordon, Rosenblum 2004). Ces capacités d'écholocalisation sont particulièrement développées chez les personnes aveugles de naissance (Schenkman, Nilsson 2010).

- Effet « cocktail party »

Pour conclure cette partie sur la localisation spatiale, notons qu'un autre résultat important de la psychoacoustique concerne la mise en évidence d'un effet « cocktail party » : grâce à sa capacité à discriminer la direction d'arrivée des sons, un auditeur est capable, au milieu d'une foule bruyante, de focaliser son attention sur une conversation en particulier (Arons 1992).

La spatialisation sonore[11] est donc particulièrement importante en robotique de téléprésence, non seulement pour que la personne se sente immergée dans l'environnement du robot, mais aussi pour renforcer l'intelligibilité des voix distantes. Elle peut être réalisée de deux manières différentes : soit en utilisant plusieurs haut-parleurs, pour pouvoir simuler différentes sources sonores (cf. son multicanal) ; soit en utilisant un enregistrement binaural, c'est-à-dire un signal

[11] Action de donner l'illusion que les sons enregistrés sont localisés dans l'espace.

stéréo qui contient les indices acoustiques évoqués précédemment. C'est cette dernière solution qui est la plus adaptée à la robotique de téléprésence, car elle nécessite très peu d'équipement du côté du pilote : un simple casque ou une paire d'écouteurs suffit.

1.4.1.2 Enregistrement binaural

L'enregistrement binaural est l'héritier direct du théâtrophone de Clément Ader. Il consiste à enregistrer le son avec une paire de microphones, puis à l'écouter au casque. Dans l'idéal, les microphones doivent être placés dans les propres oreilles de l'auditeur : ainsi, toutes les modifications acoustiques provoquées par le corps de l'auditeur sont enregistrées. Grâce à une calibration rigoureuse décrite en détails par (Møller 1992), il est alors possible de reproduire au niveau des tympans de l'auditeur quasiment les mêmes pressions que celles qu'il a éprouvé au moment de l'enregistrement.

En pratique, cette technique n'est évidemment pas envisageable, si ce n'est pour des expériences très pointues en psychoacoustique. On utilise donc des têtes artificielles, c'est-à-dire des mannequins anthropomorphes dédiés à l'enregistrement, dont la forme des pavillons est particulièrement détaillée (cf. Figure 24). La revue de (Paul 2009) présente le développement historique de ces appareils, et souligne que le principal enjeu à l'heure actuelle concerne la question de la standardisation des enregistrements binauraux. En effet, les mensurations des têtes artificielles reposent sur des mesures effectuées dans les années 60, sur des cohortes constituées principalement d'hommes adultes européens. Elles ne sont donc pas adaptées à tous les utilisateurs, et en particulier aux enfants.

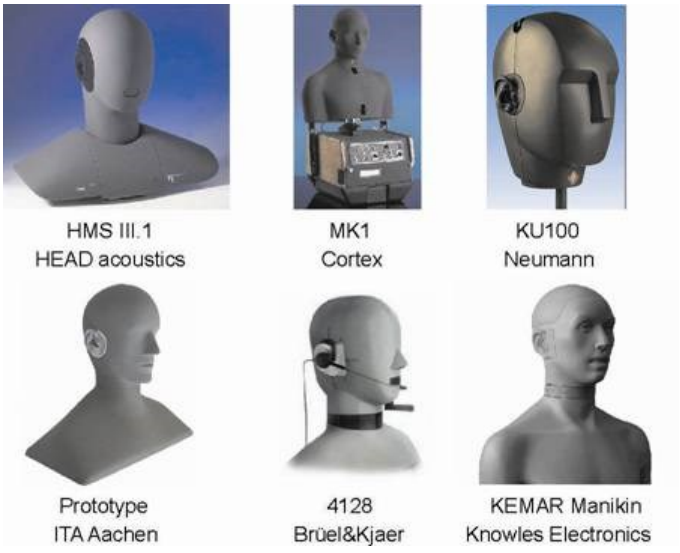


Figure 24 : Exemples de têtes artificielles (Fels, Vorlaender 2008)

1.4.1.3 Synthèse binaurale

Une autre solution pour produire du son binaural consiste à simuler les modifications acoustiques produites par le corps de l'auditeur : c'est ce qu'on appelle la synthèse binaurale. En effet, la manière dont une tête (humaine ou artificielle) modifie une onde acoustique peut être décrite à l'aide de fonctions de transfert, connues sous le sigle HRTF (Head Relative Transfer Function). Une HRTF est associée à une oreille donnée et à une direction de l'espace

(azimut + élévation). Il s'agit de la traduction en fréquences d'une HRIR (Head Relative Impulse Response), qui elle, peut être mesurée en chambre anéchoïque à l'aide de signaux artificiels. À l'aide d'un enregistrement mono et d'un couple de HRIR (une pour chaque oreille), il est possible de simuler un son binaural.

La Figure 25 présente un exemple de couple de HRTF, mesuré pour une source sonore située à 50cm, à droite de l'auditeur, à hauteur des oreilles. On constate que l'intensité acoustique parvenant à l'oreille droite est plus forte que celle parvenant à l'oreille gauche, en particulier dans les hautes fréquences. Ce résultat est conforme à la théorie de la perception spatiale détaillée précédemment.

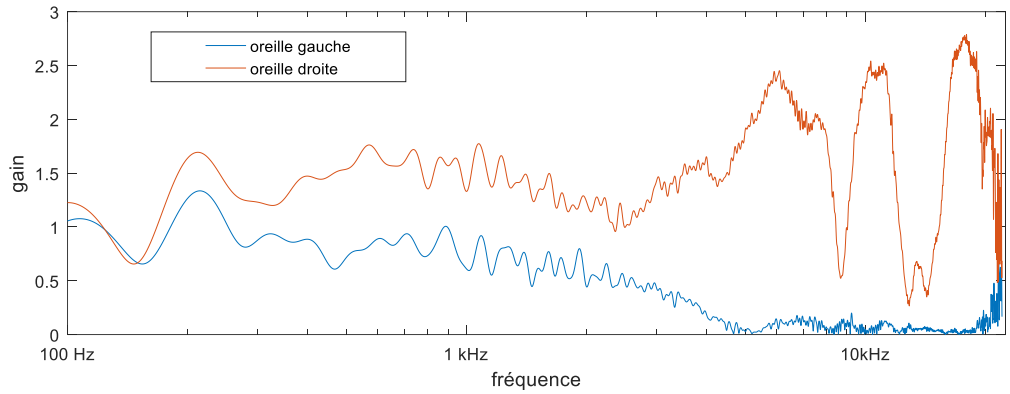


Figure 25 : Couple de HRTF mesuré à (90°, 0°, 50cm)
d'après les réponses impulsionnelles du sujet 1002 (Ircam 2002)

À l'heure actuelle, il existe plusieurs bases de données accessibles en ligne, qui regroupent à la fois des mesures effectuées avec des têtes artificielles standard et avec des êtres humains. Il y a deux limites principales à ces bases de données. Tout d'abord, il s'agit de mesures discrètes : il est impossible de mesurer de façon continue la HRIR dans toutes les directions de l'espace. Par exemple, la base de l'IRCAM comporte « seulement » 187 paires de HRIR par sujet. Un des enjeux de la synthèse binaurale à l'heure actuelle consiste donc à interpoler les HRIR en temps réel, afin de pouvoir simuler des scènes audio qui suivent les mouvements de tête des auditeurs. Un second enjeu est celui de l'individualisation : en effet, une HRIR mesurée pour un individu A, ne convient pas forcément à un individu B, car ils n'ont pas les mêmes caractéristiques physiques. Plusieurs chercheurs s'intéressent donc à la manière d'associer les propriétés des HRTF à des mesures anthropométriques, afin de pouvoir sélectionner, voire simuler, les HRIR adaptées à chaque personne. Par exemple, (Fels, Vorländer 2009) ont cherché à déterminer les paramètres qui influencent le plus les HRTF, en distinguant ceux qui concernent la forme de la tête et du torse, et ceux qui concernent la forme du pavillon de l'oreille. Dans le premier cas, ils concluent que les paramètres les plus importants sont la distance épaule – oreille, la largeur et la profondeur de la tête.

Comme une banque de HRIR se mesure en chambre anéchoïque, elle ne suffit pas à générer une scène acoustique complète. Il faut également tenir compte des modifications liées à l'acoustique du lieu simulé, modélisées par une réponse impulsionnelle de salle, ou RIR (*Room Impulse Response*). Les HRIR mesurées dans des environnements acoustiques réels sont

appelées BRIR (*Binaural Room Impulse Response*). Il existe même des OBRIR (*Oral-Binaural Room Impulse Response*), qui servent à simuler la manière dont un auditeur entendrait sa propre voix dans un environnement acoustique donné (Cabrera et al. 2009).

1.4.1.4 Écoute binaurale

Le son binaural doit être écouté à l'aide d'un casque ou d'une paire d'écouteurs, de manière à séparer le son qui parvient à l'oreille gauche, et celui qui parvient à l'oreille droite. Or, tous les casques ne sont pas équivalents, en termes de qualité sonore. En particulier, ils n'ont pas tous la même réponse en fréquences. Ainsi, la Figure 26 présente trois exemples fictifs de casques audio, de qualité bien différente. La courbe rouge correspond à un casque idéal : toutes les fréquences du signal enregistré sont reproduites parfaitement, sans être amplifiée, ou atténuée. Ce genre de casque n'existe pas en réalité : même le meilleur casque audio a une réponse en fréquences qui s'écarte de l'horizontale. Cependant, la réponse en fréquences permet de classer les casques : la courbe verte est ainsi meilleure que la courbe bleue, qui, elle, présente des pics et des creux de près de 20 dB.

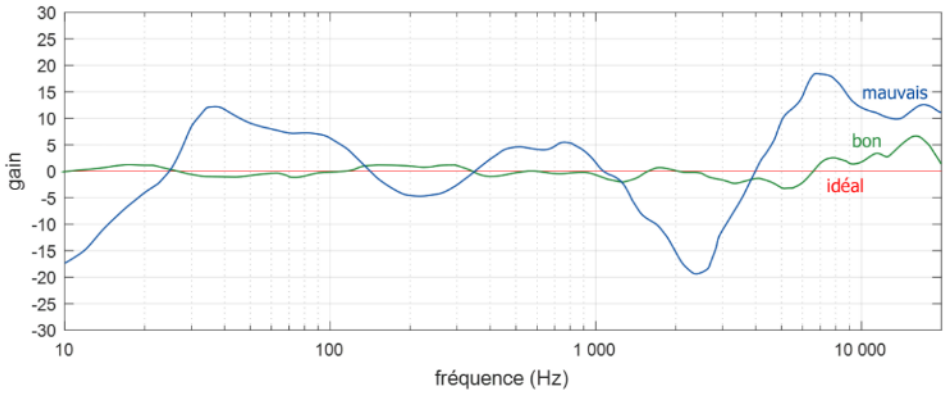


Figure 26 : Trois exemples de réponses fréquentielles (inspirés des courbes mesurées par lesnumeriques)

Ces écarts à l'horizontale se traduisent par des modifications du timbre du son. En théorie, ils peuvent être compensés à l'aide d'un égaliseur : un dispositif électronique ou un logiciel qui permet de modifier la réponse en fréquences. Cependant, il existe d'autres sources de distorsions du signal sonore, que l'égalisation ne peut pas corriger, voire risque d'empirer (ex : rapport signal sur bruit, distorsion harmonique…).

Notons que le problème se pose également pour les haut-parleurs, mais aussi pour les microphones. Lorsque le but est de transporter quelqu'un d'un lieu à l'autre, il faut donc bien choisir le matériel utilisé.

1.4.1.5 Réalisme ou fiction acoustique

Nous avons vu comment réaliser une écoute binaurale qui se rapproche le plus possible de ce que l'auditeur entendrait s'il était physiquement présent sur le lieu d'enregistrement. Cependant, les techniques de spatialisation sonores peuvent également servir à créer des « fictions acoustiques ». Ainsi, pour enregistrer un concert, on peut envisager soit de placer une tête artificielle dans le public, afin de permettre à l'auditeur d'entendre le concert comme s'il y

était ; soit d'enregistrer chaque instrument séparément, et simuler a posteriori la scène acoustique. Dans ce dernier cas, il devient possible d'amplifier certains instruments peu audibles, ou d'exacerber les indices acoustiques, ce qui permet à l'auditeur d'entendre plus clairement la scène acoustique que s'il était présent sur place.

L'évaluation de la qualité spatiale est donc un sujet complexe à traiter : s'agit-il d'évaluer la fidélité acoustique vis-à-vis de la scène initiale, ou le caractère agréable pour l'auditeur ? De plus, les critères subjectifs utilisés dans les tests de qualité audio souvent difficiles à définir (Rumsey 2002). Ainsi, on peut demander aux évaluateurs de noter la profondeur de scène, le caractère agréable, ou encore la chaleur du son. De tels tests nécessitent donc au préalable une longue phase d'apprentissage de la part des sujets, qui doivent se familiariser au vocabulaire du test, et apprendre à percevoir des variations acoustiques subtiles.

1.4.2 Bouches artificielles

L'immersion acoustique peut également être prise en compte du côté des interlocuteurs. Eux entendent la voix du pilote à travers le robot de téléprésence. Il s'agit donc de faire en sorte que le son qui sort du robot soit le plus proche possible de celui que le pilote aurait pu produire s'il était présent. Après les oreilles artificielles, nous allons donc nous intéresser aux bouches artificielles.

1.4.2.1 Technologies existantes

Les bouches artificielles sont notamment utilisées dans l'industrie de la téléphonie ou par des équipes de recherche en acoustique. Elles sont conçues pour produire un champ sonore similaire à celui produit par un être humain. En effet, la forme du corps modifie les ondes acoustiques qui arrivent aux oreilles, mais également celles qui sortent de la bouche.

Un exemple de bouche artificielle disponible dans le commerce est présenté en Figure 27. Il existe peu d'études sur le sujet, mais il semblerait que ces bouches artificielles ne reproduisent que partiellement les caractéristiques de la voix humaine (Halkosaari, Vaalgamaa 2005). En particulier, elles ne tiennent pas compte des variations dues aux changements de géométrie de la bouche au cours de la production de parole (Pollow 2015 ; Postma, Katz 2016).

Figure 27 : Exemple de bouche artificielle (Brüel & Kœjar)

À côté de ces instruments de mesures standardisés, on trouve des prototypes développés pour la recherche en acoustique. Ceux-ci ne sont pas de simples haut-parleurs aux caractéristiques

soigneusement choisis, mais des maquettes mécaniques qui imitent l'anatomie humaine. On peut citer notamment le robot Waseda Talker, visible en Figure 28. La machine possède des cordes vocales, des lèvres et même une langue articulée pour reproduire le plus fidèlement possible la parole humaine (Fukui et al. 2010).

Figure 28 : Photo du robot Waseda Talker (modèle WT-7RII)
http://www.takanishi.mech.waseda.ac.jp/top/research/voice/index.htm

En pratique, ces bouches artificielles sont très peu répandues. Un simple haut-parleur, même de taille réduite, permet d'avoir une bonne qualité sonore à condition d'être suffisamment amplifié pour que la voix porte jusqu'à ses interlocuteurs. La vraie problématique en robotique de téléprésence provient en réalité de la prise de son, que nous allons rapidement évoquer dans le paragraphe suivant.

1.4.2.2 Problématique de la prise de son

Lorsqu'on enregistre une voix, il ne suffit pas de choisir le bon matériel pour limiter au maximum les distorsions sonores : il y a également un choix à faire sur la distance d'enregistrement. En effet, une prise de son est comme une photographie : elle n'est pas une transcription fidèle de la réalité, mais un point de vue (Deshays 2006). Ainsi, placer un microphone proche de la bouche est l'équivalent d'une macrophotographie : cela permet d'isoler la voix de son environnement, pour en saisir un maximum de détails. En revanche, dès qu'on éloigne le microphone du locuteur, son timbre est modifié, l'intensité de la voix enregistrée diminue, et d'autres signaux peuvent être captés : des bruits ambiants, mais également des échos de la voix du locuteur, si l'enregistrement se fait en intérieur.

Figure 29 : Exemple de macrophotographie (Luc Viatour)

Les micros se distinguent également par leur directivité : en fonction de la direction d'arrivée du son, l'intensité captée par le microphone est plus ou moins importante. Nous en présentons quelques exemples en Figure 30. Ainsi, un microphone est qualifié d'omnidirectionnel lorsqu'il capte le son dans toutes les directions. Au contraire, un microphone cardioïde capte principalement les sons venant de l'avant ; tandis qu'un supercardioïde capte également en partie les sons venant de l'arrière.

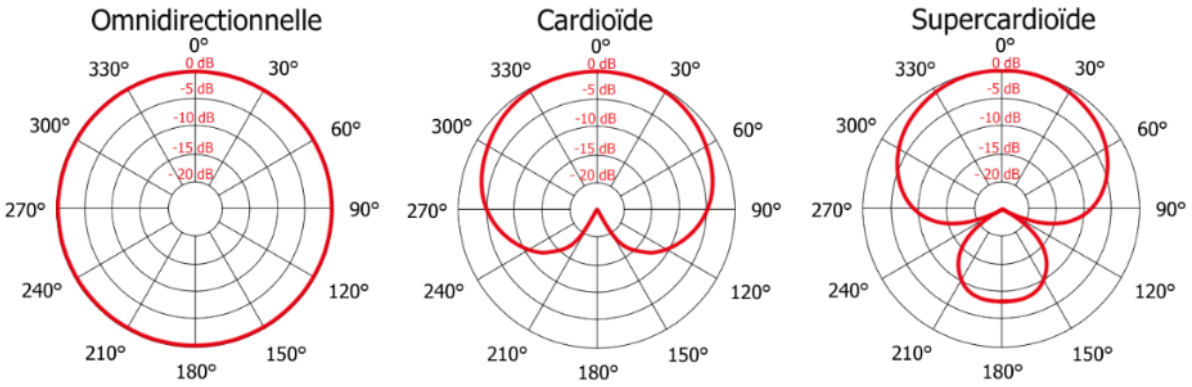


Figure 30 : Trois exemples de directivité

Le choix du microphone du pilote, et sa distance d'enregistrement, n'est presque jamais évoqué en robotique de téléprésence : le pilote utilisant généralement le microphone interne de son ordinateur. Or ces choix conditionnent la manière dont la voix va sonner dans l'environnement local.

1.4.3 Application à la robotique de téléprésence

La robotique de téléprésence hérite des recherches sur le son binaural. On y retrouve les deux approches présentées précédemment : enregistrement binaural, qui consiste à utiliser une tête artificielle pour enregistrer le son, ou synthèse binaurale, qui consiste à recréer une scène acoustique à partir de plusieurs enregistrements séparés. Dans cette partie, nous allons en voir quelques exemples, afin de préciser les avantages et inconvénients des deux méthodes.

1.4.3.1 Des têtes artificielles pilotables

Un robot de téléprésence peut très rapidement être transformé en tête artificielle : il suffit de placer deux micros, de part et d'autre de la tête du robot. Pour renforcer l'immersion, il peut être intéressant d'asservir les mouvements de tête du robot à ceux du pilote ; cela a déjà été testé par plusieurs équipes de recherche.

Ainsi, (Harrison, Mair 2007) ont conçu le système d'oreilles artificielles illustré en Figure 31. Il est constitué de deux microphones, placés de part et d'autre de la tête du robot, et entourés de pavillons reproduits par le *National Centre for Prosthetics and Orthotics* à partir des oreilles d'un des chercheurs. La tête artificielle peut tourner horizontalement à ±170°, et verticalement à ± 45°. Elle est capable d'imiter précisément les mouvements de l'auditeur, avec une erreur de moins de 0.5° lorsque celui-ci est fixe. Lorsque l'auditeur bouge, l'erreur peut-être plus importante (jusqu'à 12° dans les données de l'article). De plus, la latence entre le mouvement

du sujet et celui de la tête artificielle est non négligeable, puisqu'elle reste en moyenne autour de 150 ms, et peut monter jusqu'à 500 ms dans le pire des cas. Cette latence est donc clairement perceptible par les sujets, puisque son seuil de détection a été estimé autour de 60 ms par (Yairi et al. 2007). Les moteurs du robot sont également bruyants, puisque les chercheurs ont mesuré en chambre source une augmentation importante du bruit ambiant, entre les phases de repos (25 dB) et les phases de rotation, (44.7 dB). Notons que Harrisson et Mair s'intéressent notamment à l'influence de la distance interaurale sur les performances en localisation spatiale du téléopérateur. En effet, plus l'écart entre les deux oreilles artificielles est grand, plus il devrait être facile de localiser les sources sonores, puisque les signaux arrivant à chaque oreille sont plus contrastés.

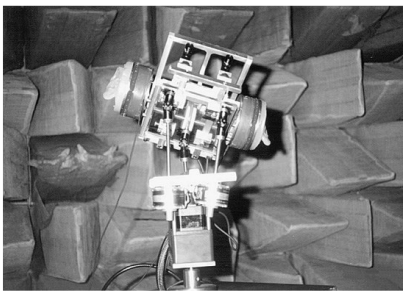


Figure 31 : Tête mécatronique développée par (Harrison, Mair 2007)

(Toshima et al. 2008) proposent un système similaire (cf. Figure 32). Leur robot a une apparence humaine très réaliste, et un degré de liberté supplémentaire, puisque sa tête peut tourner autour de l'axe avant-arrière (mouvement de roulis). Ce système est plus rapide puisqu'il peut imiter les mouvements de tête du téléopérateur dans un délai de 120 ms. Il est également moins bruyant, puisque le bruit généré pour le pilote n'est que de 24dB SPL dans la plage de fréquences 1 – 4 kHz. Ainsi, si le bruit ambiant est de 25 dB SPL, il ne monte qu'à environ 28 dB SPL dans les phases de mouvement.

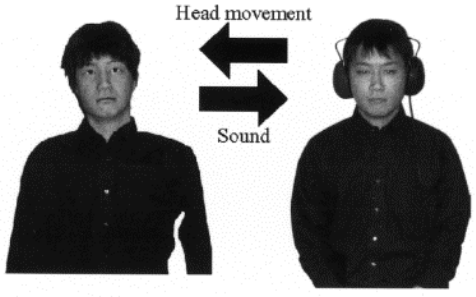


Figure 32 : Photo du robot TeleHead II et de son utilisateur (Toshima et al. 2008)

Les deux exemples précédents sont difficiles à mettre en place : il faut pouvoir imiter de façon précise l'anatomie de l'auditeur, ainsi que reproduire des mouvements de rotation complexes. (Hirahara et al. 2015) ont testé deux systèmes beaucoup plus simples : a) une tête de mannequin extrêmement simplifiée, sans nez ni pavillons ; et b) deux microphones omnidirectionnels (cf. Figure 33). Seule la rotation horizontale est reproduite, dans un délai de 120 ms. Les chercheurs ont comparé les résultats aux tests de localisation dans le cas où les oreilles étaient

immobiles, et dans le cas où elles suivaient le mouvement des auditeurs. Ils constatent que les mouvements de la tête permettent de localiser le son bien plus précisément, y compris dans le plan vertical, bien que la tête ne bouge qu'horizontalement. En outre, ces mouvements permettent de réduire la sensation de son intracrânien, très courante dans le cas d'une écoute au casque : bien qu'ils soient capables de percevoir la direction d'arrivée du son, les auditeurs peuvent avoir l'impression que celui-ci provient de l'intérieur de leur tête, et non de sources extérieures. Les chercheurs concluent donc qu'imiter précisément l'anatomie de l'auditeur a beaucoup moins d'importance si la reproduction binaurale est dynamique que si elle est statique.

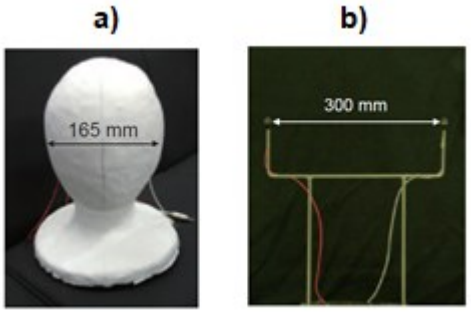


Figure 33 : Dispositifs étudiés par (Hirahara et al. 2015)

Lorsque le pilote du robot est équipé d'un casque de réalité virtuelle, un système binaural dynamique, même très simplifié, lui permet donc immédiatement de se sentir présent dans la scène acoustique. Une tête artificielle a l'avantage d'être très économe en termes de débit binaire, puisqu'elle enregistre un simple signal stéréophonique. Cependant, le suivi de mouvement peut être difficile à implémenter de façon satisfaisante : les moteurs utilisés sont bruyants, et réagissent avec un temps de retard clairement perceptible par les auditeurs, qui s'ajoute à la latence du réseau.

Dans le cas de la téléprésence ubiquïte, le suivi de mouvement n'est pas très pertinent, car si le pilote tourne la tête, il ne fait plus face à l'écran qui lui permet de voir l'environnement du robot. Un système binaural statique est donc suffisant ; ce qui ne signifie pas pour autant que la tête du robot doit être totalement fixe, simplement que ses mouvements doivent être contrôlées via l'interface du robot, comme c'est le cas pour les modèles présentés en section 1.1.2.

1.4.3.2 Des scènes acoustiques éditables

D'autres roboticiens se sont penchés sur la synthèse binaurale. Dans ce cas, le problème consiste d'abord à séparer les différentes sources sonores présentes dans l'environnement. La scène acoustique peut ensuite être resynthétisée à partir de HRTF.

La séparation de sources repose sur l'utilisation de plusieurs microphones. Il s'agit d'abord de localiser les sources sonores dans l'espace ; c'est-à-dire savoir de quelle direction vient le son. (Rascon, Meza 2017) propose une revue des méthodes de localisation de sources en robotique. Ils constatent que la majorité des travaux concernent des approches binaurales, comme le montre la Figure 34 extraite de leur article. En effet, les chercheurs tendent à imiter le système

auditif humain en utilisant deux microphones, situés de part et d'autre du robot, et entourés de pavillons artificiels. Il s'agit également d'une volonté de limiter au maximum le nombre de microphones utilisés ainsi que le coût algorithmique, car le système de localisation doit pouvoir être embarqué sur un robot aux ressources limitées. Dans des cas extrêmes, des systèmes à 1 microphone peuvent être imaginées si les sources sonores sont connues à l'avance. Les solutions à 8 microphones sont également très populaires, car elles représentent un bon compromis en termes de précision et de quantité de données traitées.

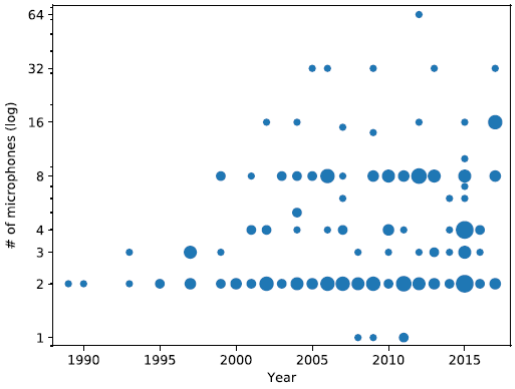


Figure 34 : Evolution du nombre de microphones utilisés dans la littérature sur la localisation de source sonore – extrait de (Rascon, Meza 2017)

Une fois les sources sonores localisées, il est possible de traiter les enregistrements, pour en extraire des signaux séparés : un pour chaque source sonore. Ces signaux sont ensuite filtrés par des HRTF et combinés pour créer un signal binaural. Les HRTF utilisées sont choisies en fonction de la direction des sources sonores, et de l'orientation de la tête locuteur. Le système de (Liu et al. 2019) représenté en Figure 35 illustre clairement ces différentes étapes.

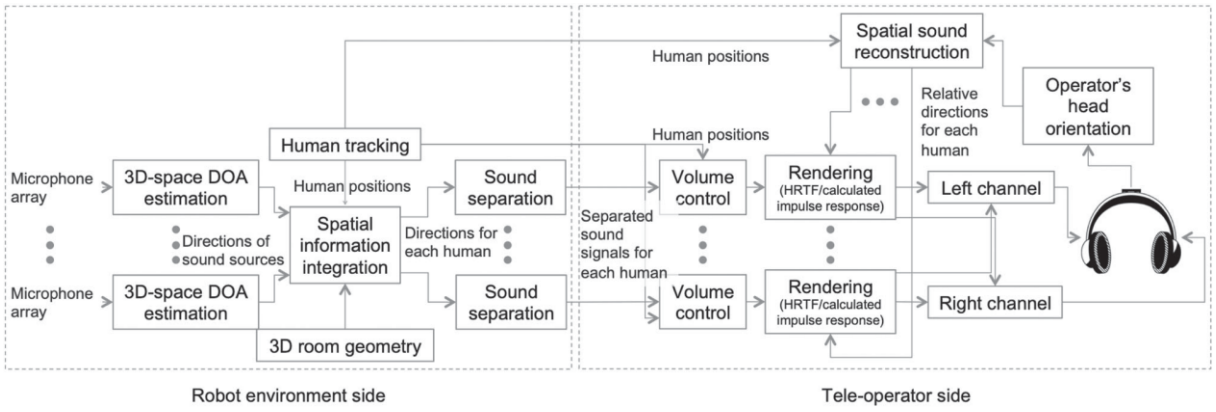


Figure 35 : Système de téléprésence proposé par (Liu et al. 2019)

La synthèse binaurale est donc envisageable en robotique de téléprésence, bien qu'elle soit plus difficile à implémenter qu'un simple enregistrement binaural. Les contraintes matérielles sont importantes, car le système de localisation doit pouvoir être embarqué sur un robot aux ressources de calculs et d'énergie limitées. En particulier, comme le robot se trouve le plus souvent en intérieur, il peut être difficile de séparer la source sonore directe de ses échos.

Cependant, le net avantage de la synthèse binaurale est qu'elle permet un suivi de tête silencieux, car le robot n'a pas besoin de tourner la tête : seul le mixage des différents canaux acoustiques évolue en temps réel avec les mouvements de tête de l'auditeur. De plus, l'utilisation de techniques de synthèse binaurale permet d'envisager de rendre la scène acoustique « éditable » : par exemple, en contrôlant le volume de chaque source sonore de façon indépendante, comme l'ont proposé (Liu et al. 2015).

1.4.4 Résumé

Il existe déjà des technologies grand public pour l'immersion acoustique. Certaines sont faciles à mettre en œuvre en robotique de téléprésence : en particulier, il suffit de deux micros pour réaliser un enregistrement binaural, qui permette au pilote de localiser la direction d'arrivée des sons. Les bouches artificielles sont beaucoup plus anecdotiques, et limitées principalement aux études acoustiques.

1.5 Conclusion

Ce premier chapitre nous a permis de présenter un état de l'Art sur la robotique de téléprésence. Les robots de téléprésence sont des systèmes de télécommunication, conçus pour transmettre la présence d'une personne grâce à des technologies immersives. Les modèles les plus sophistiqués sont encore à l'étude en laboratoire, cependant il en existe déjà un certain nombre disponible dans le commerce, et utilisés au quotidien, en particulier pour le télétravail. Tous suivent à quelques exceptions près le même modèle : un écran qui porte l'image de l'utilisateur, posé sur une base mobile.

S'ils permettent effectivement d'augmenter la présence de leur utilisateur, ces robots ne transmettent pas encore parfaitement son « toucher social » : il ne peut pas percevoir et agir dans l'environnement social distant comme s'il y était. Ce manque de dextérité à manipuler les signaux sociaux à distance peut être lourd de conséquences, en particulier dans le cas où la téléprésence est conçue comme un moyen de maintenir le lien social.

Chapitre 2 :
TOUCHER VOCAL

Le premier chapitre consacré aux robots de téléprésence nous a permis d'annoncer la problématique de la thèse : Comment assurer un **« toucher vocal »** en conditions de **téléprésence ubiquïte** ? C'est-à-dire, comment permettre au pilote du robot de téléprésence de réaliser une proprioception à distance de ses signaux socio-affectifs ? En particulier, comment lui permettre de contrôler sa **« portée vocale »**, selon les compétences qu'il a acquise au cours de ses interactions *in situ* ?

Ce second chapitre, consacré plus précisément au toucher vocal, nous permettra d'aborder les notions théoriques et techniques nécessaires pour répondre à cette problématique. Nous commencerons par présenter quelques **notions fondamentales** pour l'étude de la parole : en particulier, nous verrons les principales grandeurs physiques permettant de décrire les signaux vocaux. Puis, nous chercherons à définir la notion de **« portée vocale »**, qui n'existe pas encore réellement en tant que telle dans la littérature. Cette portée vocale, qui permet au locuteur de contrôler à qui il s'adresse, est également perceptible par les auditeurs présents, indépendamment de l'intensité qui parvient à leurs oreilles : ils peuvent donc en déduire des informations concernant le locuteur. En particulier, nous présenterons plusieurs études en psychoacoustique qui montrent que la **perception de la distance** peut être influencée par la perception de la portée vocale.

2.1 La parole : signaux physiques, intrinsèquement sociaux ?

Cette première section constitue une introduction très générale au signal de parole. Elle va nous permettre de définir le vocabulaire et les connaissances basiques qui nous serviront dans la suite pour étudier le toucher vocal. Nous commencerons par présenter rapidement la physiologie de la parole, puis nous nous intéresserons aux principales grandeurs utilisées pour caractériser les signaux vocaux.

2.1.1 Physiologie de la parole

*La production de parole met en jeu plusieurs organes, qui constituent « **l'appareil phonatoire** », représenté en* Figure 36.

Tout d'abord, les **poumons** (1) constituent une soufflerie permettant de générer le flux d'air nécessaire à la phonation.

Le **larynx** (2) fait la jonction entre les voies respiratoires inférieures et supérieures. Il abrite la **glotte**, un sas ouvert pendant la respiration pour laisser passer le flux d'air, et fermé au moment de la déglutition pour bloquer le passage des aliments vers les poumons et les guider vers l'œsophage. L'ouverture et la fermeture de la glotte passe par la contraction des deux **plis vocaux** (ou cordes vocales). Ce sont ces plis vocaux qui permettent la vocalisation : lorsqu'ils sont tendus, il est possible de les faire vibrer à l'aide d'un flux d'air sortant des poumons. Les sons produits sont alors qualifiés de **voisés**.

La **sphère pharyngo-bucco-nasale** (3), abrite tous les organes nécessaires à la mastication, et permet la respiration. Elle sert également de résonateur, permettant de moduler les vibrations et le flux d'air issus du larynx. Le mouvement des différents articulateurs (langue, dents et lèvres) peut même produire des sons non pulmonaires, tels que les clics utilisés dans certaines langues.

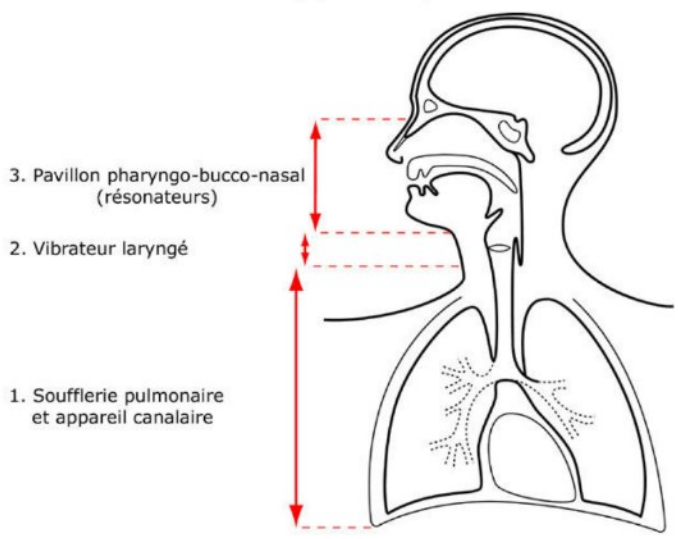


Figure 36 : Schéma de l'appareil phonatoire (Gabriel 2018)

Enfin, le centre de contrôle de l'appareil phonatoire est le **cerveau**. L'étude de patients souffrant d'aphasie à la suite de lésions cérébrales a permis de mettre en évidence l'importance de deux régions cérébrales en particulier : l'**aire de Broca**, associée à la production de parole, et l'**aire de Wernicke**, associée à sa compréhension (cf. Figure 37). Ces aires sont reliées entre elles par un ensemble de fibres nerveuses, le **faisceau arqué**, ainsi qu'aux cortex moteur, visuel et auditif (Démonet [sans date]). L'hypothèse dominante à l'heure actuelle considère qu'il existerait des **connexions neuronales directes** entre le cortex moteur et les organes de production de la parole. Ce sont ces connexions directes qui ferait la spécificité du langage humain, et non le conduit vocal lui-même (Fitch 2018). En effet, elles permettraient une motricité fine du larynx, que ne possèdent pas les autres primates. Pour les muscles du visage et de la langue en revanche, l'existence de ces connexions neuronales directes est attestée à la fois chez l'humain et les primates.

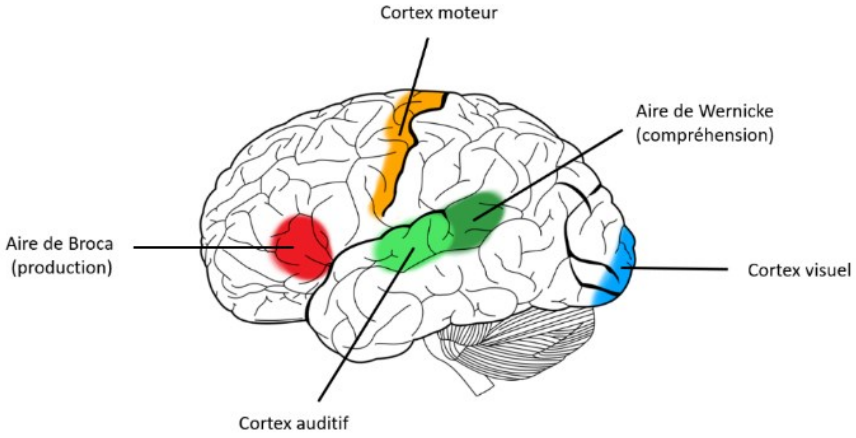


Figure 37 : Principales régions cérébrales impliquées dans la parole

À ces structures qui permettent la perception acoustique et la production de parole, il faut ajouter également les **aires émotionnelles** du cerveau. Les notes de cours de (Berthoz 2003) présentent un historique des théories sur la physiologie des émotions, et décrivent les principales régions impliquées ainsi que leurs fonctions émotionnelles et cognitives.

2.1.2 Couplage perception-action

À travers ce court résumé de la physiologie de la parole, nous avons vu que les aires cérébrales dédiées à la production et et à la perception de la parole sont reliées entre elles. Ce **couplage entre perception et action** (aussi dit sensori-moteur) est très étudié, et fait l'objet de nombreuses théories et modèles. À titre d'exemple, on peut citer l'effet Lombard, sur lequel nous reviendrons plus en détails au chapitre 5 : il s'agit d'une adaptation de la production de parole engendrée par la perception d'un bruit.

Un des grands enjeux scientifiques à l'heure actuelle consiste à proposer des modèles qui intègrent également une dimension **socio-affective** et **empathique**. En particulier, l'existence des **neurones miroirs** a été mise en évidence en constatant que le fait d'observer une personne

effectuer un geste active dans le cerveau de l'observateur les zones du cerveau qui lui permettraient de réaliser ce geste (Rizzolatti, Craighero 2004).

Notre thèse s'inscrit dans cette démarche, puisque nous nous intéressons à la proprioception et à l'inter-proprioception **sociale** de cette boucle sensori-motrice. En effet, l'objectif n'est pas seulement de permettre au pilote du robot de téléprésence de contrôler sa portée vocale, mais également de faire en sorte que ses interlocuteurs sachent qu'il la contrôle (cf. Figure 38).

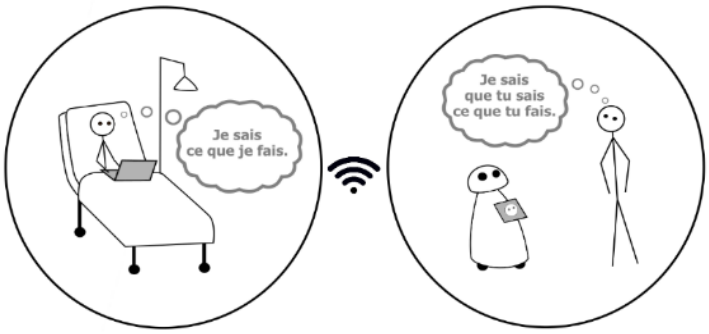


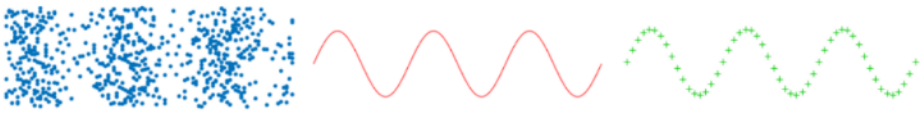
Figure 38 : Proprioception et inter-proprioception

2.1.3 Descriptions des signaux de parole

Nous allons à présent définir quelques notions utiles pour décrire les signaux de parole.

2.1.3.1 Intensité

Les signaux de parole sont des signaux acoustiques, et peuvent donc être étudiés sous trois formes différentes : acoustique, électrique, et numérique. Des mesures d'intensité spécifiques ont donc été définies par chaque discipline : les principales sont résumées dans la Figure 39, et décrites de façon détaillée en Annexe B. Soulignons qu'une partie de ces mesures ont été conçues spécifiquement pour estimer la sonie, c'est-à-dire la sensation auditive produite chez un être humain. En l'absence de précision, toutes les mesures en dB présentées dans nos résultats de thèse seront des mesures numériques, obtenues à l'aide de la méthode présentée en Annexe B.



	Acoustique	Electrique	Numérique
Mesure du niveau d'intensité	dB(SPL)	dB(V), dB(u)	dB(Praat), dB(FS)
Mesure de la sonie	dB(A), dB(B), dB(C)		LUFS

Figure 39 : Récapitulatif des principales mesures de l'intensité sonore

Mesurer une intensité à un instant donné n'a pas de sens, puisque l'amplitude de l'onde évolue en permanence. L'intensité est donc toujours mesurée sur une fenêtre temporelle, dont la taille peut aller de quelques millisecondes, à plusieurs dizaines de secondes. Une petite fenêtre permet de visualiser l'enveloppe temporelle du signal ; tandis qu'une fenêtre plus large fournit une mesure d'intensité moyenne, qui permet de déterminer si la personne parle fort ou doucement (cf. Figure 40).

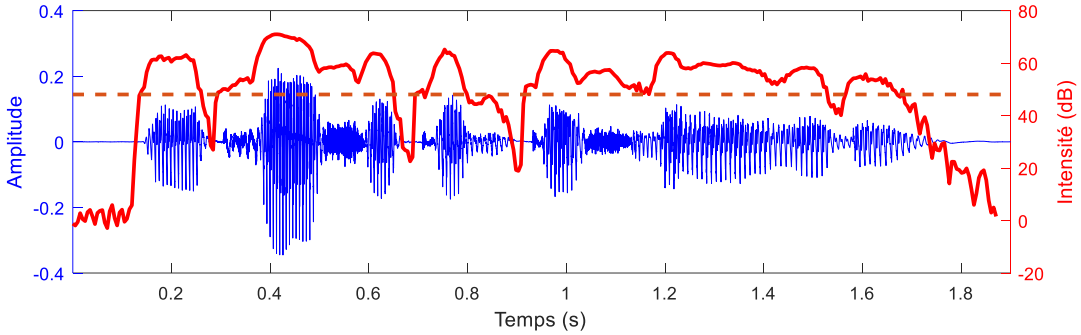


Figure 40 : Enregistrement de parole extrait du spectacle Aporia
(« Montrez ce que vous cachez derrière le dos »)
La ligne en pointillés représente l'intensité moyenne du signal.

Soulignons qu'une mesure d'intensité n'est pas absolue, mais relative, puisqu'elle dépend de la distance à laquelle se fait l'enregistrement. Ce qui caractérise une source sonore, ce n'est pas son intensité I, mais sa **puissance acoustique** P, qui se répartit sur une surface sphérique de plus en plus grande à mesure que l'on s'éloigne de la source sonore : $I = \frac{P}{2\,\pi\,r^2}$. En théorie, cette diminution d'intensité est donc de $10\,\log_{10}(4) \approx 6$ dB par doublement de la distance. S'il n'est pas possible de mesurer P directement, sa valeur peut être estimée à partir d'une mesure de I à une distance de référence : le plus souvent 1m, par convention.

La puissance acoustique d'un locuteur dépend directement du débit d'air en sortie de ses poumons. Par analogie, en soufflant très fort dans une flûte, on produit un son plus fort qu'en jouant normalement[12]. Parler fort nécessite donc un effort physique ; c'est pourquoi, les études portant sur des voix de différentes puissances font souvent référence à la notion d'**effort vocal**. Cependant, cette notion est mal définie et peu pertinente lorsqu'on considère toutes les productions vocales possibles : en particulier, dans le cas du chuchotement, il faut un effort physique important pour produire de la parole avec une puissance acoustique faible. Dans cette thèse, nous préférerons donc l'expression **force de voix**, proposée par Jean-Sylvain Liénard, afin de distinguer la puissance acoustique produite par un locuteur, de l'intensité perçue par les sujets. La Figure 41 représente une échelle indicative de différentes forces de voix.

[12] En réalité, la physique de la flûte est bien plus complexe que cela, et ce n'est pas en soufflant le plus fort possible qu'on obtient le son le plus fort (Terrien 2014).

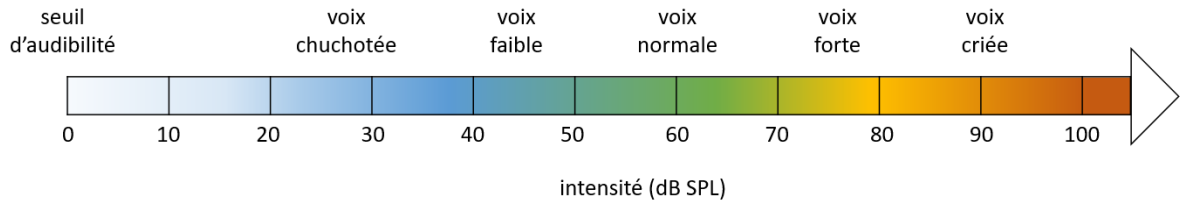


Figure 41 : Échelle approximative de la force de voix, mesurée à 1m

Des données plus détaillées sont disponibles dans l'étude de (Pearsons et al. 1977). Elle montre notamment que la force de voix varie en fonction de l'environnement acoustique et des locuteurs : en particulier, les femmes participant à l'étude parlaient en moyenne moins fort que les hommes (-7 dB pour les voix criées).

2.1.3.2 Durée

Un enregistrement de parole peut être découpé en sous-parties : par exemple, syllabe par syllabe, comme en Figure 42. L'analyse de ce découpage, et de la durée de chaque section, peut nous fournir des informations intéressantes. Ainsi, il est possible d'accéder au **débit de parole,** c'est-à-dire le nombre de syllabes par seconde. Plus le locuteur parle vite, plus le débit de parole est élevé. Il faut également penser à mesurer les silences et les pauses, qui font partie intégrante du signal de parole.

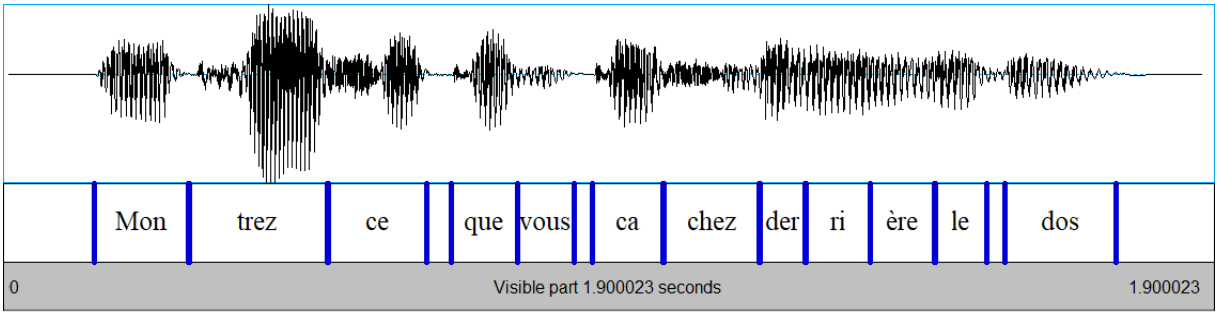


Figure 42 : Découpage syllabique d'un enregistrement du spectacle Aporia

2.1.3.3 Fréquence fondamentale

La **fréquence fondamentale**, aussi appelée **pitch** et notée généralement **F0**, est la fréquence de vibrations des plis vocaux. Elle détermine la hauteur du son, c'est-à-dire, le fait qu'il soit perçu comme aigu ou grave. Cette fréquence fondamentale dépend d'abord de caractéristiques anatomiques : l'épaisseur et la longueur des plis vocaux, qui varient d'un individu à l'autre. Ainsi, la voix d'un enfant de deux ans s'élève à environ 400 Hz. En grandissant et sous l'effet des hormones sexuelles à la puberté, la voix devient plus grave. À l'âge adulte, les fréquences fondamentales communément admises sont donc de 200-250 Hz pour les femmes, et 100-150 Hz pour les hommes (Tubach 1989). Il s'agit là uniquement de valeurs moyennes, obtenues en parole standard : en réalité, la palette vocale des locuteurs est beaucoup plus étendue, puisqu'ils sont capables de contrôler la tension de leurs cordes vocales pour modifier leur pitch. À travers

la classification des différentes tessitures des voix chantées, on apprend par exemple que la fréquence fondamentale de la voix la plus grave (basse) peut descendre à 60 Hz, tandis que celle de la voix la plus aigüe (soprano) peut monter jusqu'à 1200 Hz. À titre indicatif, la Figure 43 représente un dessin anatomique de la glotte, telle qu'elle peut être observée au cours d'un examen médical à l'aide d'un endoscope, placé dans la gorge du patient.

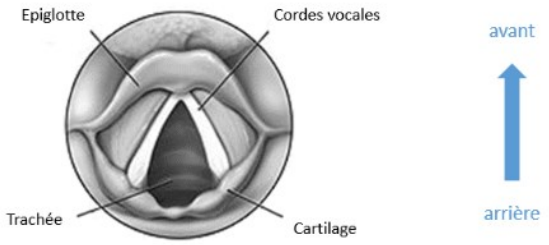


Figure 43 : Schéma anatomique en vue du dessus de la glotte (National Cancer Institute)

2.1.3.4 Timbre et qualité de voix

Cette fréquence fondamentale s'accompagne de différentes harmoniques, multiples de la fréquence fondamentale. L'amplitude de ces harmoniques définit le **timbre** de la voix. En musique, le timbre est ce qui distingue deux instruments jouant la même note : leur fréquence fondamentale est identique, mais le poids des différentes harmoniques varient d'un instrument à l'autre. En phonétique, le timbre est lié à la taille et à la **forme du conduit vocal**, qui amplifie certaines harmoniques, et en atténue d'autres.

Le mouvement des articulateurs (langue, dents et lèvres) permet de produire les différents sons de la langue. En particulier, les voyelles sont caractérisées par des **formants**, notés F1, F2 etc., qui correspondent aux maximums d'intensité du spectre fréquentiel (cf. Figure 44).

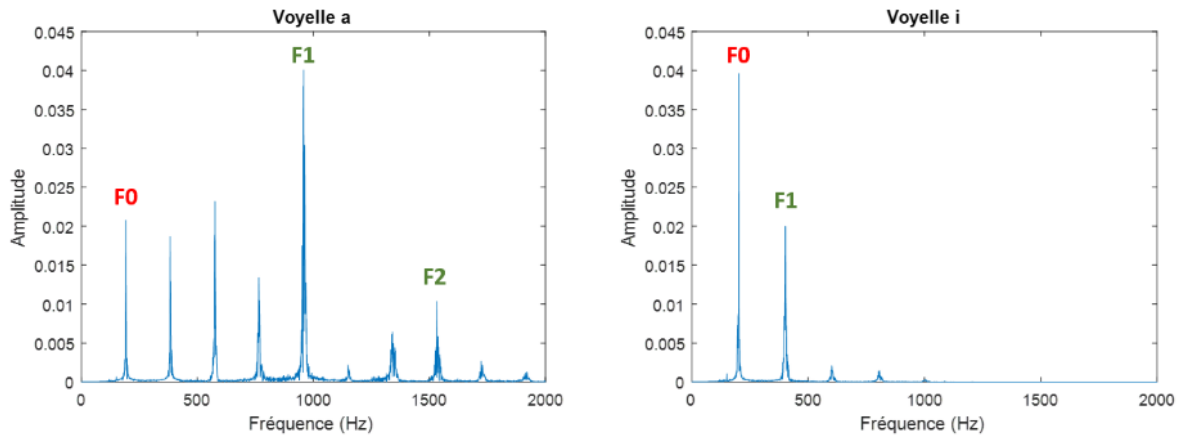


Figure 44 : Comparaison des spectres de deux voyelles prononcées à la même fréquence fondamentale.
Rouge : fréquence fondamentale | Vert : Formants

Il est également possible de faire varier l'**ouverture de la glotte**, en jouant sur les différents muscles du larynx. Les plis vocaux peuvent ainsi vibrer selon différents modes, dont (Laver 1980) a proposé une taxonomie. Cette taxonomie est décrite en détails et complétée par des études plus récentes dans la thèse de (Audibert 2008). Quelques exemples de configuration de la glotte sont présentés en Figure 45.

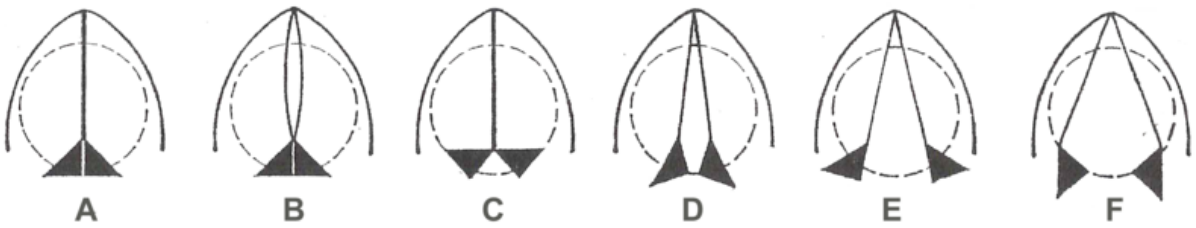


Figure 45 : Schéma simplifié des différentes configurations de la glotte (Wikipédia : article anglais sur la phonation)

La position A correspond à une fermeture totale de la glotte, qui peut être observée au moment de la déglutition pour empêcher les aliments de passer dans les poumons. Pendant la phonation normale, dite **modale**, la glotte alterne rapidement entre les positions A et B, sous l'effet du souffle émis par les poumons, qui force l'ouverture de la glotte. Il s'agit du mode de vibration optimal : toute l'énergie des poumons sert à faire vibrer la glotte. Cependant, il existe également des modes de vibration sous-optimaux. Ainsi, en position C (voix **murmurée**) et D (voix ***breathy***), une partie du souffle s'échappe, car la glotte est en permanence entrouverte. La voix perd donc en intensité, et un bruit de friction s'y ajoute. Enfin, la position E correspond à un état de repos, ou de respiration ; tandis que la position F peut être observée en cas de respiration profonde.

La **tension** des plis vocaux engendre également des différences de timbre : ainsi, la voix est qualifiée de ***lax***, lorsque la tension musculaire est faible, ou ***tense*** dans le cas contraire. En particulier, une voix ***harsh*** est particulièrement tendue, ce qui engendre des vibrations irrégulières en durée et en amplitude. Les voix ***fry*** ou ***creaky***, sont également des voix tendues, mais associée à une fréquence fondamentale basse.

Le timbre de la voix peut être altéré par certaines **pathologies vocales** ou par le **vieillissement**, qui peut provoquer soit un affinement, soit un épaississement des cordes vocales, ou encore une réduction de la puissance pulmonaire (Hagen et al. 1996). Ainsi, les grands fumeurs, tels que Jeanne Moreau ou Charles Gainsbourg, sont connus pour leurs voix rauques, abimées par la chaleur et les toxines dégagées par la fumée de cigarette.

Du point de vue perceptif, les différents **modes de phonation** engendrent différentes **qualités de voix**, dont la perception a été étudiée, en particulier en phonostylistique. On peut citer notamment les travaux de (Fónagy 1983) et ceux de (Léon 1993), qui se sont intéressés aux impressions produites sur les auditeurs par l'écoute de différentes voix. La socio phonétique a également mis en évidence des différences de timbre liées au **genre**, qui ne peuvent pas s'expliquer uniquement d'un point de vue physiologique. Ainsi, selon (Pépiot 2013), l'identification du genre repose sur des représentations mentales, des attentes sur ce qui est une

voix féminine ou une voix masculine, propres à la langue et à la culture. Notamment, les voix *breathy,* sont plutôt associées aux femmes, et les voix *fry* ou *creaky* aux hommes. Les études de (Arnold, 2015), qui s'intéresse à la perception des voix de femmes transgenres, montrent également l'importance des fréquences de résonnance, et donc de l'articulation. Parler d'une voix féminine ou masculine relèverait donc de la performance de genre, et donc d'un apprentissage.

Ces recherches concernant l'apparence vocale trouvent aujourd'hui des applications en **robotique sociale** : ainsi, (Walters et al. 2008) et (Moore 2017) ont étudié l'impact de différentes esthétiques vocales sur l'interaction avec un robot, et en particulier sur la distance à laquelle leurs sujets s'en approchaient.

En pratique, la **modélisation acoustique** de la qualité de voix reste cependant un problème à l'heure actuelle, car il n'est pas possible d'avoir accès à la forme de la glotte, ou à la tension musculaire des plis vocaux à partir d'un simple enregistrement acoustique. Il existe donc trois méthodes principales pour évaluer la qualité de voix. La première consiste à faire entendre la voix à un **auditeur expert** (phonéticien), capable d'identifier les modifications anatomiques à partir du son produit. La seconde méthode consiste à mesurer certaines caractéristiques du signal, qui dans la littérature ont déjà été **corrélées** avec la qualité de voix. La troisième méthode consiste à utiliser des **appareils de mesure** supplémentaires : par exemple, l'électroglottographe est un collier constitué de deux électrodes placées au niveau du larynx, qui sert à mesurer les variations d'impédance provoquées par l'ouverture et la fermeture de la glotte.

2.1.4 Étude de la prosodie

En réalité, les grandeurs physiques qui permettent de décrire la parole, et que nous avons séparées en quatre catégories distinctes (intensité, durée, fréquence fondamentale et timbre), ne sont pas entièrement indépendantes les unes des autres, comme nous allons le voir à travers quelques exemples.

Ainsi, (Bernardoni 2012) note que seuls les chanteurs entraînés sont capables de séparer leur **fréquence fondamentale** de leur **intensité de voix**, et ce uniquement lorsqu'ils chantent. En règle générale, tout locuteur a tendance à parler plus aigu lorsqu'il parle plus fort (Gramming et al. 1988). En effet, si le débit d'air issu des poumons est plus élevé, il faut une tension plus forte des plis vocaux pour les maintenir en position ; ce qui conduit à une élévation de la fréquence fondamentale. Ce décalage vers les aigus n'affecte pas uniquement la fréquence fondamentale, mais le spectre de la voix dans son ensemble, donc son **timbre**. Ainsi, (Liénard 2018) montre que le maximum et le barycentre du spectre moyen long terme (SMLT)[13] se décale vers les aigus lorsque la force de voix augmente. C'est également sur le spectre que (Turk et al. 2005) se basent pour manipuler l'effort vocal de phrases prononcées dans une langue imaginaire avec trois niveaux différents (doux, moyen et fort). (Fux 2012), lui, conclut à la fin de sa thèse que la piste la plus pertinente pour transformer une voix normale en voix

[13] Il s'agit d'une mesure de l'intensité d'un signal par bandes de fréquences.

criée (sans modifier son intensité) consiste à modifier l'évolution de sa fréquence fondamentale au cours du temps : c'est-à-dire à jouer à la fois sur la F0 et sur la **durée** du signal.

Par ailleurs, ces informations acoustiques **ne sont jamais perçues séparément** par les auditeurs. En particulier, (Traunmüller 1994) estime que si la reconnaissance des syllabes passait uniquement par la valeur des formants, il serait impossible de comprendre à la fois une voix d'enfant et une voix d'homme, tant leurs caractéristiques diffèrent. À la place, il propose une théorie de la modulation : la parole serait le résultat de gestes vocaux conventionnels, produits à partir de perturbations d'un signal porteur, qui lui dépend des caractéristiques biologiques du locuteur. Pour pouvoir décoder le signal de parole, l'auditeur devrait donc passer par une phase de démodulation, afin d'extraire ces gestes vocaux. C'est cette étape de démodulation qui manque cruellement aux systèmes actuels de traitement automatique de la langue, qui, à moins d'y avoir été entraîné spécifiquement, ne parviennent pas à reconnaître des détails qui sembleraient pourtant évident à un auditeur : ainsi, ils ne peuvent pas reconnaître qu'une voix a un accent, ou qu'elle est ironique. Pour apprendre à reconnaître la parole, ou à l'imiter, ces systèmes doivent se baser sur de vastes corpus d'exemples, sélectionnés et annotés par des humains. Cette méthode se révèle particulièrement inefficace : ainsi, (Dupoux 2018) estime qu'un système de reconnaissance de la parole à l'état de l'art tel que Deep Speech 2 nécessite 10 000 heures de parole annotée, et les probabilités d'apparition de quelques milliards de mots ; au contraire, un enfant maya de 4 ans par exemple, n'a eu accès qu'à 14 fois moins de temps de parole, et 240 fois moins de mots pour apprendre à parler sa langue maternelle.

La notion qui permet de faire le lien entre intensité, durée, fréquence fondamentale et qualité de voix est la **prosodie**. Elle recouvre un domaine extrêmement vaste, englobant tous les phénomènes dits suprasegmentaux, c'est-à-dire qui ne peuvent pas être réduits à l'étude isolée des sons de la langue (ce qui est le domaine de la **phonétique**). La prosodie concerne ainsi l'étude de l'intonation, du rythme, ou encore des accents. Elle est porteuse d'informations **linguistiques** (par exemple, les questions du français se terminent par une prosodie montante) ; mais également **paralinguistiques**. Ainsi, c'est à travers la prosodie que s'expriment les émotions (Gobl 2003 ; Johnstone, Scherer 1999 ; Scherer 1995). La prosodie sert également à marquer les attitudes, tels que la politesse et l'autorité, ou encore le rôle social et l'appartenance à un groupe.

Une des principales hypothèses de notre équipe de recherche est que la prosodie nourrit une « **glu socio-affective** » (Aubergé 2015 ; Sasa 2018). Décrire l'état de la glu qui relie deux agents, c'est décrire l'état de leur relation à un moment donné. Les informations paralinguistiques n'ont donc rien d'anecdotiques, mais constituent au contraire l'essentiel de la communication. Paul Watzlawick, un des fondateurs de l'école de Palo Alto, le résumait ainsi :

> *Un cinquième, peut-être, de toute communication humaine sert à l'échange de l'information, tandis que le reste est dévolu à l'interminable processus de définition, confirmation, rejet et redéfinition de la nature de nos relations avec les autres. (Watzlawick , traduit par (Winkin 1981), p. 240)*

2.1.5 Résumé

Dans cette première partie, nous avons présenté plusieurs notions fondamentales pour l'étude de la parole. Bien que contraint par des caractéristiques anatomiques, le signal de parole se révèle extrêmement variable d'un locuteur à l'autre, et même chez un même locuteur. Il contient une quantité d'informations considérable, concernant à la fois l'identité du locuteur, ses socio-affects, ce qu'il veut dire, et comment il le dit. L'étude de la prosodie est une manière d'aborder cette variabilité, pour en extraire les détails qui nous intéressent. Dans notre cas, il s'agit des détails porteurs d'information socio-affective, qui constituent le toucher vocal.

2.2 La portée vocale

Dans cette thèse, nous nous intéressons à un aspect du toucher social en particulier : celui qui concerne la gestion de la portée vocale. Nous définissons la portée vocale comme la distance, ou plus largement, la zone de l'espace, dans laquelle doit se trouver l'interlocuteur pour comprendre ce que le locuteur veut dire. Cette portée vocale est très difficile à mettre en équations, mais nous chercherons tout de même à en tracer les contours, en l'encadrant d'une marge supérieure et d'une marge inférieure.

2.2.1 Définition par l'intensité

En français, être « à portée de voix », signifie être audible par une autre personne. L'expression anglaise suit la logique inverse : *within earshot*, signifie littéralement « à portée d'oreille ». La portée vocale est donc liée directement à la force de voix : plus une personne parle fort, plus sa portée vocale est grande puisqu'il est possible de l'entendre à grande distance.

Une première définition de la portée vocale est donc de considérer qu'il s'agit de la zone de l'espace dans laquelle la voix d'un locuteur est **audible**. En définissant un seuil d'audibilité, il serait alors possible de calculer directement la portée vocale, à partir de la force de voix. Imaginons par exemple un locuteur parlant au milieu d'un champ. En théorie, selon la loi de propagation du son, si l'intensité acoustique mesurée à 1 m est de 60 dB SPL, elle ne serait plus que de 30 dB SPL à 32 mètres, ce qui correspond au niveau d'un chuchotement. On pourrait alors considérer que la portée vocale est tout simplement une sphère de 32 mètres de rayon autour du locuteur. En fixant le seuil d'audibilité à 0 dB SPL, ce qui est la limite théorique à laquelle un son est audible pour une personne normo-entendante, on obtiendrait une sphère de 1 km de rayon (cf. Figure 46).

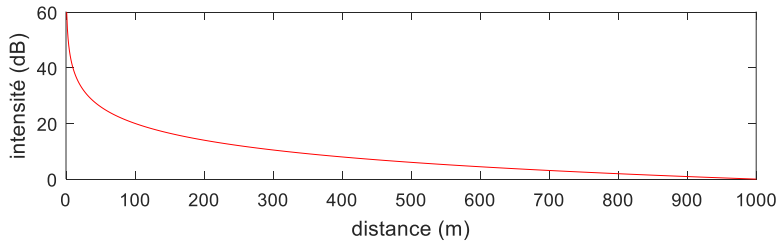


Figure 46 : Diminution théorique de l'intensité d'une source sonore en champ libre

Même en supposant une absence totale de bruit, cette première définition paraît absurde : il est très rare de parler avec une personne située à plusieurs dizaines, voire centaines de mètres. Commençons par raffiner ce modèle acoustique. En première approximation, nous avons choisi une répartition sphérique de la puissance acoustique. En réalité, la voix humaine possède une **directivité** : son énergie ne se répartit pas de façon homogène dans toutes les directions. La Figure 39 montre ainsi des mesures d'intensité obtenus par (Chu, Warnock 2002) à partir d'un locuteur masculin. On constate que la majeure partie de l'énergie acoustique est dirigée vers l'avant. En outre, lorsque ce locuteur parle avec une voix chuchotée, son intensité est beaucoup plus atténuée derrière la tête, que lorsqu'il parle avec une voix modale, ou criée. Sa voix chuchotée serait donc particulièrement peu audible pour un auditeur qui ne se tiendrait pas face à lui. Cependant, ce résultat est contesté par (Monson et al. 2012), qui le considère comme un cas particulier : en étudiant un groupe de 15 sujets (dont 8 femmes), ils concluent que la directivité de la voix humaine ne varie que faiblement en fonction du genre et de la force de voix. Nous illustrons malgré tout ce paragraphe à l'aide des schémas de (Chu, Warnock 2002), car leur lecture est beaucoup plus aisée que celle des graphiques de (Monson et al. 2012), et permet de bien se représenter la notion de directivité.

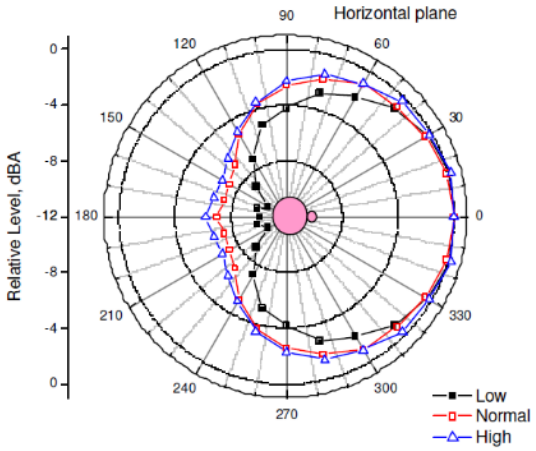


Figure 47 : Directivité de la voix d'un locuteur masculin parlant avec trois forces de voix (Chu, Warnock 2002)

Pour définir la portée vocale d'un point de vue purement acoustique, il faudrait donc tenir compte de la directivité de la voix humaine, et la représenter sous une forme de cardioïde, ou de sphère aplatie. On peut ensuite étendre la définition de la portée vocale, en considérant qu'il ne s'agit pas simplement de pouvoir entendre que la personne parle, mais également de savoir ce qu'elle dit.

2.2.2 Définition par l'intelligibilité

Une seconde définition de la portée vocale est de considérer qu'il s'agit de la zone de l'espace dans laquelle la voix d'un locuteur est **intelligible**. Or, l'intelligibilité de la parole ne se limite pas à l'intensité. Pour l'évaluer, il faut faire appel à des cohortes de sujets, et vérifier s'ils sont capables dans des conditions données de comprendre ce qui est dit. Les résultats obtenus dépendent fortement du protocole choisi.

Pour limiter les tests à la compréhension acoustique seule, on utilise généralement des mots isolés, des logatomes (mots sans signification) ou des phrases dénuées de sens (Lambourg 2002) : le test d'intelligibilité consiste alors à retranscrire exactement les mots entendus. Cependant, il peut être intéressant d'évaluer l'intelligibilité dans des situations de communication verbale réalistes, c'est-à-dire lorsque l'auditeur a accès à des éléments de contexte (grammatical, sémantique…) qui facilitent sa compréhension. La thèse de (Fontan 2012) a ainsi portée sur le développement de tests d'intelligibilité destinés à évaluer les capacités de communication de personnes atteintes de troubles pathologiques de la production de la parole. Ces personnes peuvent obtenir des scores très faibles aux tests d'intelligibilité classiques, tout en étant bien comprises pendant leurs interactions quotidiennes.

En pratique, ces tests perceptifs sont extrêmement coûteux à mettre en place, en particulier s'ils tiennent compte de la durée d'apprentissage des sujets : ainsi, pour des corpus de grande taille (ex : 20 listes de 50 éléments de test), les performances des auditeurs ne se stabilisent qu'au bout de plusieurs heures d'entraînement (Lambourg 2002). Un certain nombre d'indices prédictifs de l'intelligibilité ont donc été développés, inspiré du modèle télégraphique de Shannon (cf. (Fontan 2012) pour un historique détaillé). Ces critères permettent d'évaluer non pas l'intelligibilité des locuteurs, mais celle du canal de communication. Par exemple, l'indice STI (Speech Transmission Index) consiste à mesurer le rapport signal sur bruit (RSB) entre un signal artificiel[14] et le bruit de fond (Steeneken, Houtgast 2014). Cet indice compris entre 0 et 1 apparaît dans plusieurs textes de normalisation (ex : ISO9921 et IEC 60268-16).

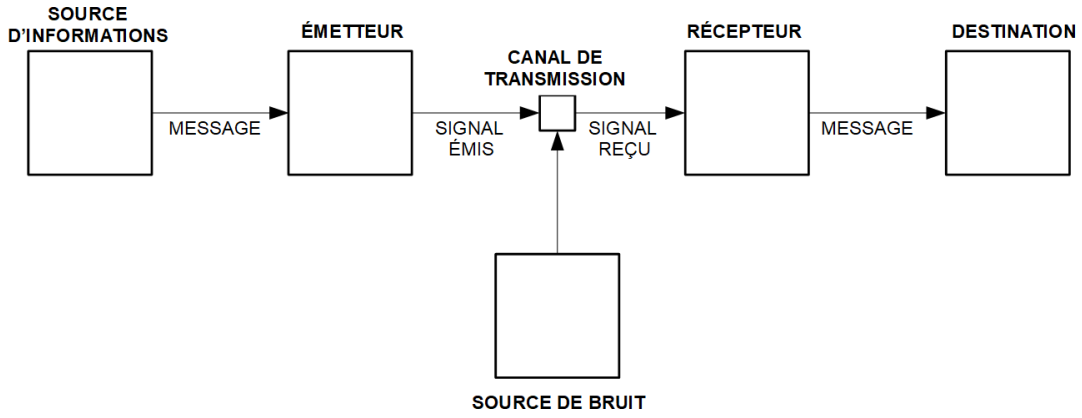


Figure 48 : Modèle télégraphique de la communication (Shannon 1948)
Si ce modèle a longtemps été « la » référence en sciences humaines et sociales sous l'appellation « modèle de Shannon et Weaver » (Picard 1992), son objet initial n'est pas vraiment la parole. En effet, il s'agit avant tout d'un schéma illustratif qui permet à Shannon de poser les bases de sa théorie de l'information, une théorie mathématique qui va servir à développer les technologies de la télécommunication. La question qu'il se pose est de savoir comment transmettre un message, de quelque nature que ce soit, d'un point A à un point B, sachant qu'au cours de la transmission, ce message peut être altéré. La « source de bruit » qu'il évoque n'est donc pas un bruit acoustique, mais représente toutes les altérations subies par le signal. Par exemple, dans le cas d'une copie d'ADN en biologie, le bruit correspondrait à des mutations génétiques. Shannon s'intéresse donc aux probabilités d'erreur, et cherche à coder l'information de manière à la rendre redondante pour qu'une fois arrivé à destination le message puisse être préservé.

[14] Plusieurs signaux artificiels de caractéristiques temporelles et fréquentielles variées sont utilisés, et la somme de leurs RSB pondérée pour obtenir le STI.

L'intelligibilité évoquée ci-dessus est une intelligibilité purement linguistique ; mais, on pourrait également imaginer une intelligibilité socio-affective. En effet, un cri d'énervement n'a pas le même impact, selon que les interlocuteurs sont proches ou distants l'un de l'autre. De même, il semble difficile de réconforter une personne à grande distance, puisque plus la distance est grande, plus il est difficile d'avoir une voix douce. La notion d'intelligibilité devrait donc prendre en compte non seulement ce que dit le locuteur mot pour mot, mais également ce qu'il veut dire, et ne pas dire.

2.2.3 Définition par le confort auditif

Avoir une bonne portée vocale consiste donc à parler suffisamment fort pour se faire entendre, en tenant compte du lieu et de ses caractéristiques acoustiques. Mais il s'agit également de ne pas parler trop fort, ce qui serait une source de fatigue pour le locuteur et d'inconfort pour l'auditeur.

Cette notion de **confort acoustique** a surtout été étudiée en audiologie, en particulier dans le but de régler des appareils d'aide auditive (Punch et al. 2004). Par exemple, (Hawley et al. 2017) ont demandé à leurs sujets de classer des sons de différentes intensité selon 7 catégories allant de « très faible » à « désagréablement fort ». Ils ont utilisé plusieurs stimuli : tons purs de 500 Hz, 2000 Hz et 4000 Hz, et mots isolés. Pour les mots isolés, un pallier est franchi tous les 10 dB environ, et le niveau d'intensité « confortable » s'élève aux alentours de 60 dB HL. Cette valeur moyenne de 60 dB, qui est généralement admise comme référence pour la parole sur les échelles de bruit, a également été obtenue par (Sato et al. 2007). Leur étude se déroulait avec un haut-parleur en chambre anéchoïque. Plusieurs mots-clés étaient diffusés à différents niveaux d'intensité et en simulant une réverbération plus ou moins importante.

2.2.4 Résumé

À partir des définitions précédentes, on peut tracer une représentation schématique de la portée vocale (cf. Figure 49). Cette portée vocale est une zone de l'espace délimitée par deux frontières : une frontière basse, qui correspond à la distance en dessous de laquelle la voix est trop forte pour être confortable, et une frontière haute, qui correspond à la distance au-delà de laquelle la voix est trop faible pour être intelligible.

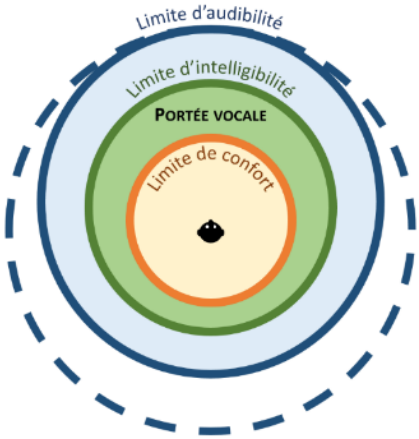


Figure 49 : Schéma illustratif de la portée vocale

Ces frontières théoriques sont extrêmement difficiles à tracer en pratique, puisqu'elles dépendent à la fois du locuteur, de l'auditeur et de l'environnement acoustique. En outre, ces frontières sont dynamiques, puisque l'environnement acoustique évolue en permanence. Pourtant, dans des conditions in situ, un locuteur est capable d'adapter sa portée de voix pour se faire comprendre en fonction de ce qu'il veut dire, et de à qui il s'adresse. C'est en cela que la portée vocale fait partie du toucher vocal : une portée vocale adaptée dans un contexte, ne l'est pas forcément dans un autre. Cette adaptation à l'environnement est très clairement illustrée par les travaux des urbanistes (Augoyard et al. 1985), qui constatent que le bruit n'est pas seulement un « grand séparateur social » empêchant la communication vocale, mais qu'il assure également trois fonctions essentielles : « gérer la portée, la destination et les limites de l'acte de communiquer ».

Dans cette thèse, nous ne chercherons pas à modéliser mathématiquement la portée vocale, mais à mettre en évidence le lien entre portée vocale, socio-affects et proxémie : en particulier, une voix douce, c'est une voix faible, et une proximité physique et sociale. La question fondamentale est de savoir s'il est possible de conserver les caractéristiques d'une telle voix en téléprésence. Pour répondre à cette question, nous avons choisi un chemin détourné, en étudiant la manière dont la distance physique est perçue en téléprésence (chapitre 4) et la manière dont la portée vocale peut être modifiée par la situation d'ubiquïté (chapitre 5).

2.3 Perception acoustique de l'espace social

Dans le chapitre 1, nous nous sommes intéressés à localisation des sources sonores, pour pouvoir ensuite comprendre les technologies permettant de créer une immersion acoustique en téléprésence. Dans cette partie, nous nous intéressons spécifiquement aux études qui concernent la localisation spatiale de locuteurs, ou de sources de parole enregistrées. En particulier, nous verrons que la perception de la distance physique qui sépare l'auditeur du locuteur est étroitement liée à la perception de sa force de voix.

2.3.1 Familiarité des signaux de parole et définition de la distance

Rappelons que pour percevoir la distance, un auditeur a à disposition deux indices principaux : l'intensité et le rapport d'énergie directe / réverbérée (cf. §1.4.1.1). L'intensité est un indice relatif, utilisable uniquement lorsque la source se déplace, ou si l'auditeur a une idée a priori de son intensité ; tandis que le rapport d'énergie directe / réverbérée est un indice absolu, mais dépendant de l'environnement acoustique.

Or, la plupart des études en psychoacoustique utilisent des signaux artificiels, ce qui permet de contrôler finement leurs propriétés spectrales et temporelles. De plus, les premières recherches concernant la perception de la distance se sont faites dans la lignée des études sur la localisation azimutale : c'est-à-dire en chambre anéchoïque, ou dans des environnements où le rapport d'énergie directe / réverbérée est négligeable. En conséquence, il était presque impossible pour les sujets de percevoir la distance des sources sonores. C'est le constat réalisé par (Coleman 1962) : dans son étude, les auditeurs doivent déterminer parmi une colonne de 14 haut-parleurs lequel émet un bruit blanc aléatoire d'une seconde et de 65 dB (mesuré à 30 cm face au haut-

parleur). L'expérience se déroulant sur un lac gelé recouvert de neige, aucun écho ne parvient aux auditeurs, et ils ne peuvent se fier qu'à des variations de timbre ou d'intensité pour tenter de deviner la distance relative des nouveaux stimuli par rapport aux précédents. Coleman conclut donc qu'un auditeur n'est pas capable d'estimer la distance d'une source sonore avec laquelle il n'est pas familiarisé.

D'autres résultats suggèrent que la familiarité des stimuli permettrait d'améliorer la perception de la distance. Ainsi, (McGregor et al. 1985) ou encore (Wisniewski et al. 2012) constatent que leurs sujets parviennent mieux à repérer la distance d'enregistrements de parole, lorsque ceux-ci sont joués dans le sens normal, plutôt que lorsqu'ils sont joués à l'envers. Soulignons que dans leurs cas, ce n'est pas la distance de la source sonore qui varie, mais la distance d'enregistrement : les phrases ont été enregistrées à 2 m et à 30 m du locuteur, puis leur intensité a été normalisée.

À partir de ces premiers exemples, une ambiguïté apparaît dans ce qu'on entend par perception de la distance d'une source de parole : s'agit-il de la **distance d'écoute** entre la source sonore et l'auditeur, ou de la **distance d'enregistrement** entre le locuteur et le microphone ? Par ailleurs, les signaux de parole sont des signaux sociaux. Ils incorporent donc une **distance de communication**, c'est-à-dire la distance physique entre le locuteur et son interlocuteur, réel ou imaginaire. Cette distance de communication est perceptible par les auditeurs, et peut influencer leur perception de la distance, comme nous allons le voir dans la suite.

2.3.2 Perception de la force de voix

Lorsque la distance de communication augmente, le locuteur est contraint de parler plus fort, car l'intensité qui arrive aux oreilles de l'auditeur diminue avec l'éloignement. Autrement dit, il doit augmenter sa portée vocale, en augmentant sa force de voix. Or, celle-ci ne se réduit pas à une augmentation d'intensité, puisqu'elle s'accompagne de modifications de la fréquence fondamentale, du spectre ou encore de la prosodie (cf. $ 2.1.4). Ces indices acoustiques peuvent être perçus par les auditeurs.

Ainsi, (Fux, Zimpfer 2009) ont montré que leurs sujets étaient capables de classer des enregistrements normalisés en fonction de leur force de voix. Le test consistait à écouter des paires de phrases courtes prononcées avec deux niveaux d'intensité différents, mais égalisées de manière à avoir la même sonie. Après avoir entendu chaque paire d'enregistrements, les sujets devaient répondre à la question suivante : Lequel des deux sons présentés paraît provenir de plus près ? Nous proposons une réinterprétation de leurs résultats dans le Tableau 2. Les pourcentages en diagonales correspondent aux taux de bonne reconnaissance : par exemple, pour la première case, lorsque la voix 1 était modale, elle était perçue plus proche que toutes les autres voix dans 66 % des cas. Les autres pourcentages correspondent aux taux d'erreur : ainsi, lorsque la voix 1 était modale, et la voix 2 forte, les sujets ont répondu que la seconde voix était plus proche dans 27.5% des cas. Cette étude montre qu'il est possible de deviner la portée vocale d'un locuteur, même en l'absence d'information sur l'intensité produite. Notons que l'ordre de présentation semble avoir une importance : ici, l'erreur n'est que de 5% pour la paire voix forte – voix modale, et monte à 27.5 % pour la paire voix modale – voix forte.

Tableau 2 : Taux d'erreurs et de bonne reconnaissance de l'étude de (Fux, Zimpfer 2009), cités dans (Fux 2012)

		Force de la voix 1			
		modale	forte	criée	hurlée
Force de la voix 2	modale	**66 %**	5 %	2 %	2.5 %
	forte	27.5 %	**77.5 %**	10 %	2 %
	criée	6.5 %	15.5 %	**86.5 %**	8 %
	hurlée	0 %	2 %	1.5 %	**87.5 %**

2.3.2.1 Influence de la distance de communication sur la perception de la distance

Non seulement la distance de communication est perceptible par les auditeurs, mais elle est également connue pour influencer la perception de la distance. Ainsi, les travaux de (Gardner 1969), (Mershon 1997), (Brungart, Scott 2001) et (Eriksson, Traunmüller 2002) rapportent tous une augmentation de la distance perçue avec la force de voix. Ces articles sont résumés dans les pages suivantes dans l'ordre chronologique, qui correspond également à une progression logique du protocole le plus simple au plus complexe.

L'étude de (Gardner 1969) a été menée dans une chambre anéchoïque des laboratoires Bell. Les stimuli utilisés étaient à la fois des enregistrements, et de la parole prononcée en directe par un locuteur. Ce sont les résultats concernant ce deuxième type de signaux qui nous intéressent. En effet, le locuteur devait produire quatre forces de voix : chuchotée, faible, conversationnelle et criée. Il se tenait à une distance de 90 cm, 3 m, 6 m ou 9 m du sujet. Les sujets avaient les yeux bandés, et un bruit masquant était diffusé pendant les déplacements du locuteur. Ils avaient pour consigne de ne pas se fier à l'intensité de la voix pour estimer la distance. Les résultats de cette expérience sont présentés en Figure 50. La diagonale en pointillés fins qui apparaît sur les figures de gauche représente le résultat obtenu dans un cas idéal où toutes les réponses données par les sujets sont justes. La ligne en pointillés gras représente la moyenne des réponses des sujets. Comme le choix de réponses était restreint aux quatre positions du locuteur, la distance maximale était fixée à 4, l'auteur a considéré que cela biaisait les résultats de l'expérience : la ligne continue représente les résultats qu'il prévoit en corrigeant ce biais. On constate que la distance est correctement estimée pour les forces de voix faibles et conversationnelles. En revanche, la voix chuchotée est perçue systématiquement plus proche qu'elle ne l'est réellement : la distance maximale est estimée à 3 m en moyenne, soit 3 fois moins que la distance réelle du locuteur. Au contraire, la voix criée est perçue systématiquement plus éloignée qu'elle ne l'est réellement. C'est la seule voix pour laquelle les sujets se trompent pour la distance la plus proche (90 cm), puisque 50 % d'entre eux perçoivent le locuteur à 3 m.

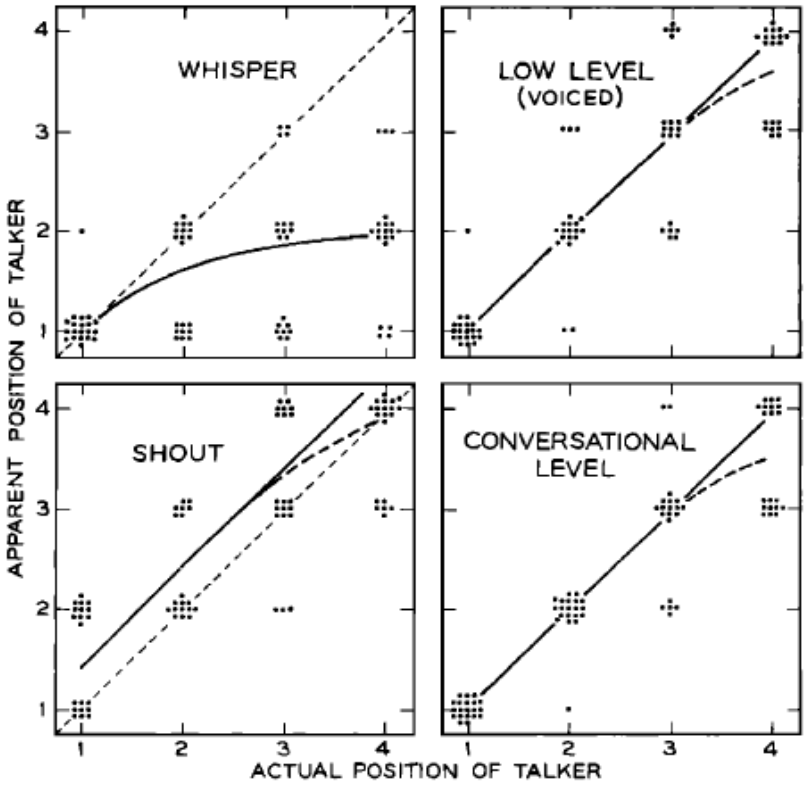


Figure 50 : Figure extraite de l'article de (Gardner 1969)
nombre de sujets : 10

Dans l'étude de (Mershon 1997), les sujets de trouvaient dans une pièce de 7,2 × 7,2 m aux sols et aux murs couverts de matériaux absorbants. Ils devaient estimer à l'aveugle la position d'un haut-parleur, situé à 2,5 m d'eux. Pour répondre, ils disposaient d'un levier situé à côté d'eux, qu'ils devaient pointer vers la source sonore (cf. Figure 51). Cette méthode de mesure est particulièrement intéressante, puisqu'elle ne nécessite pas de demander directement aux sujets d'estimer la distance en mètres, ce qui est très difficile. En outre, c'est une mesure continue, et non discrète, ce qui empêche l'apparition du biais observé par Gardner. Une simple formule trigonométrique permet de convertir l'angle du levier en distance. Les stimuli étaient des enregistrements de la phrase (« How does my voice seem ? »), prononcée avec trois forces de voix différentes par des locuteurs masculins et féminins. Les enregistrements avaient été égalisés, cependant le cri restait en moyenne plus fort que la voix chuchotée et la voix conversationnelle. L'auteur a considéré que ce déséquilibre n'était pas problématique, car il s'opposait à la tendance a priori selon laquelle un stimulus plus fort paraît plus proche. Les résultats de cette expérience rejoignent ceux de Gardner : la distance perçue est significativement plus faible pour les voix chuchotées que pour les voix conversationnelles, et la voix criée est perçue beaucoup plus éloignée en moyenne.

Ces résultats ont été confirmés par une expérience similaire, passée par 192 sujets (Philbeck, Mershon 2002). Les sujets avaient pour consigne de juger la distance « apparente » de la source sonore, plutôt que la distance « objective ». Cette fois, les chercheurs s'intéressaient à l'ordre de présentation des stimuli, pour déterminer si celui-ci pouvait influencer leurs estimations de

distance. Les sujets étaient séparés en six groupes, correspondant aux six ordres de présentations possibles pour trois stimuli. Chaque participant entendait quatre stimuli : le quatrième étant une répétition du premier stimulus entendu. Les conclusions de l'étude sont que l'ordre de présentation a peu d'influence sur la distance perçue. Ainsi, les auditeurs se baseraient avant tout sur leur « familiarité » avec les signaux de parole, plutôt que sur des comparaisons à court terme entre les différents stimuli entendus au cours de l'expérience.

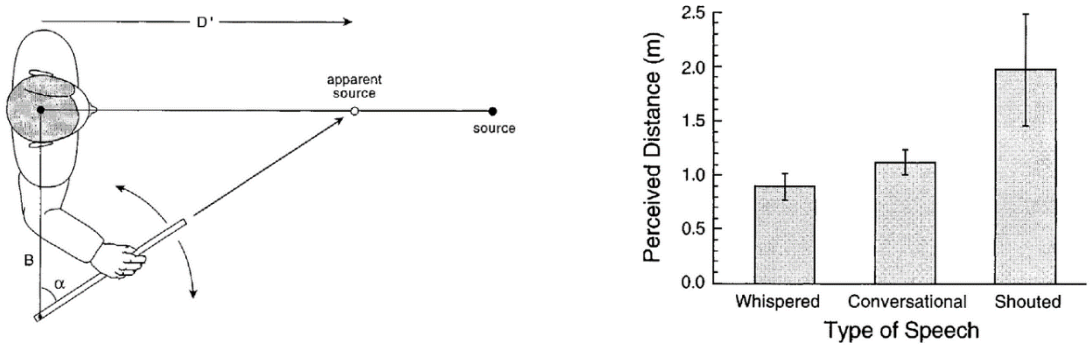


Figure 51 : Figures extraites de l'article de (Mershon 1997)
A gauche : schéma du dispositif de repérage de la distance
A droite : distance perçue en fonction du type de parole
nombre de sujets : 72

(Brungart, Scott 2001), eux, ce sont intéressés à la fois à l'intensité perçue par les sujets (*presentation level*), et à celle produite par le locuteur (*production level*). Ils ont réalisé trois expériences au casque avec des enregistrements de parole. L'expérience se déroulait dans un vaste champ d'herbe, et neuf marqueurs étaient placés (de 25 cm à 64 m) pour permettre aux sujets d'indiquer la distance qu'ils percevaient. Ils avaient le droit d'indiquer des distances intermédiaires : par exemple, 0,5 pour un son situé à mi-chemin entre eux et le marqueur numéro 1. Leurs réponses ont par la suite été converties en mètres par les expérimentateurs. Nous nous intéressons en particulier aux résultats de l'expérience II (cf. Figure 52). Pour cette expérience, les stimuli étaient des expressions (« Over here », « Threat », « Warning »), prononcées par six locuteurs (3 hommes et 3 femmes), enregistrés en chambre anéchoïques. Ces stimuli étaient traités à l'aide de HRTF mesurées pour chaque sujet pour une source située à 1 m. Leur intensité était manipulée à chaque distance, de sorte que l'intensité au niveau du sujet soit de 36, 42, 48, 54, 60, 66, 75, 78, 84, 90, ou 96 dB SPL. Les résultats obtenus montrent à nouveau que la distance perçue dépend de la force de voix du locuteur : quel que soit le niveau d'intensité perçu par le sujet, la voix chuchotée est perçue plus proche que la voix faible, elle-même perçue plus proche que la voix forte. En particulier, les auditeurs imaginent des distances cohérentes avec le niveau de production : moins de 1m pour une voix chuchotée, entre 1 et 2 m pour une voix faible, et plus de 4 m pour une voix forte. Leurs estimations étaient également cohérentes avec l'intensité perçue, puisque plus le son était fort, plus la source leur semblait proche.

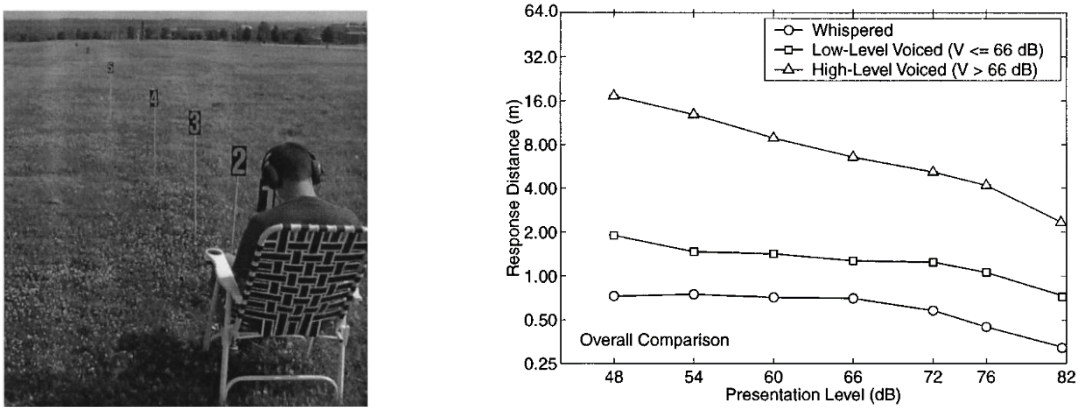


Figure 52 : Figures extraites de l'article de (Brungart et Scott 2001)
A gauche : photo du dispositif expérimental
A droite : distance perçue en fonction du type de voix et de l'intensité de la source
nombre de sujets : 6

Enfin, (Eriksson, Traunmüller 2002) ont voulu savoir si leurs auditeurs étaient capables de distinguer distance de communication et distance d'écoute. Leurs deux expériences ont eu lieu en chambre anéchoïque. La source de son était placée au plafond, à 3,5 m des sujets. La luminosité dans la chambre anéchoïque était réduite pour que les sujets ne puissent pas voir le haut-parleur. Les sujets devaient donner une estimation de la distance qu'il percevait à partir d'une liste de choix possibles. Pour faire varier la distance de communication, les stimuli utilisés étaient des enregistrements de voyelles voisées ou chuchotées, produites par deux locuteurs (un homme et une femme) s'adressant à un expérimentateur situé à 1,5 m, 6 m et 24 m (voyelles voisées) et à 0,375 m, 1,5 m et 6 m (voyelles chuchotées). Les stimuli n'étaient pas normalisés, de manière à conserver la différence d'intensité relative entre les différentes voyelles, et les différentes forces de voix. Pour simuler une variation de la distance d'écoute, les enregistrements étaient présentés avec différents niveaux d'amplification : -12, -6 et 0 dB (voyelles voisées) et -6, 0 et +6 dB (voyelles chuchotées). Dans la première expérience, les sujets étaient répartis en deux groupes : le groupe A devait estimer la distance de communication, le groupe B devait estimer la distance d'écoute. Les 12 sujets du groupe A ont perçu une augmentation de la distance de communication, lorsque la force de voix augmentait et lorsque le niveau d'amplification augmentait. Cette augmentation était plus marquée pour les voyelles voisées, que pour les voyelles chuchotées. Au contraire, les 12 sujets du groupe B ont perçu une diminution de la distance d'écoute lorsque la force de voix augmentait et lorsque le niveau d'amplification augmentait. La diminution était plus importante pour les voyelles voisées que pour les voyelles chuchotées. Pour modéliser ces résultats, les chercheurs ont calculé le logarithme en base 2 des variables étudiées, et utilisé un modèle linéaire afin d'estimer le poids de chaque variable. Leurs analyses pour les voyelles voisées sont traduites dans le Tableau 3.

Tableau 3 : Résumé des analyses présentées dans l'article de (Eriksson, Traunmüller 2002) pour les voyelles voisées

	Groupe A Distance de communication $r^2 = 0.80$		**Groupe B** Distance d'écoute $r^2 = 0.74$	
	Poids	p-value	Poids	p-value
Force de voix + 6 dB	**+0,903**	< 0,001	- 0,227	< 0,001
Intensité intrinsèque de la voyelle +6 dB	+0,467	< 0,01	- 0,308	< 0,05
Niveau d'amplification + 6 dB	+0,308	< 0,001	**- 0,730**	< 0,001
Locuteur		< 0,001		non significatif

Ces résultats témoignent d'une grande confusion de la part des auditeurs. En effet, pour estimer la distance de communication, ils devraient se fier uniquement à la force de voix ; tandis que pour estimer la distance d'écoute, ils devraient se fier uniquement au niveau d'amplification. Or, ces deux indices influencent les réponses des deux groupes. Les sujets sont également influencés par les différences d'intensité intrinsèques entre les différentes voyelles.

Les chercheurs ont donc conçu une seconde expérience, dans laquelle les sujets étaient conscients de l'existence de ces deux distances. Après chaque stimulus, il devait évaluer les deux distances. Les sujets étaient séparés en 2 groupes de 20 : l'un évaluait d'abord la distance de communication ; l'autre évaluait d'abord la distance d'écoute. Ces deux groupes étaient à nouveau subdivisés en groupe de 10 : l'un entendait les stimuli produits par la locutrice ; l'autre entendait les stimuli du locuteur. Les performances des sujets étaient similaires à celles obtenues au cours de la première expérience. Ainsi, même en étant conscients de l'existence des deux distances, ils n'étaient pas capables de les distinguer. L'ordre de présentation avait une importance, car les sujets avaient plus de mal à estimer la distance d'écoute lorsqu'elle était demandée après la distance de communication ; autrement dit, ils se souvenaient plus facilement de la force de voix avec laquelle la voyelle était produite, que de son intensité.

Quoique très différentes, ces quatre études arrivent donc à la même conclusion : pour déterminer la distance d'un locuteur, nous nous fions à sa force de voix, c'est-à-dire à la distance de communication.

2.3.2.2 Savoir à qui s'adresse le locuteur

La distance de communication fournit une autre information importante : elle permet de deviner à qui s'adresse le locuteur, dans le cas où plusieurs interlocuteurs potentiels sont présents. Ainsi, si la distance de communication est plus grande que la distance d'écoute, l'auditeur peut comprendre que ce n'est pas à lui que l'on s'adresse, ou du moins pas seulement.

Par ailleurs, comme la voix possède une directivité, il est possible de deviner l'angle vers lequel un locuteur est orienté. (Kato et al. 2010) ont ainsi réalisé une expérience en chambre

anéchoïque, qui montre que des sujets sont capables d'estimer à l'aveugle approximativement l'angle vers lequel un locuteur présent devant eux est orienté. Les résultats ne variaient pas de façon significative en fonction de la distance testée (1,2 m ou 2,4 m). (Edlund et al. 2012), eux, ont choisi de réaliser leur expérience dans des conditions plus réalistes : cinq participants aux yeux bandés étaient assis en arc de cercle dans les canapés d'un espace de détente. Un expérimentateur leur demandait plusieurs fois d'indiquer vers qui il était tourné, et les participants devaient répondre par geste. Les résultats montrent que les sujets parvenaient bien à reconnaître l'orientation de l'expérimentateur, malgré la présence de bruits provenant du couloir et des bureaux adjacents. En outre, certaines orientations étaient plus faciles à reconnaître que d'autres, ce qui suggère que les sujets étaient capables de s'aider de leur connaissance de l'environnement acoustique : ainsi, la position la mieux reconnue était celle dans laquelle l'expérimentateur était tourné vers une fenêtre, surface réfléchissante.

2.3.3 Résumé

Les premières expériences sur la perception de la distance ont mis en évidence l'importance de la familiarité de l'auditeur avec les signaux utilisés : un auditeur est incapable de percevoir la distance d'une source sonore nouvelle, dans un environnement inconnu. En revanche, on observe des résultats intéressants dans le cas où les stimuli utilisés sont des voix humaines : en effet, ces signaux sont extrêmement riches et porteurs d'informations sociales que l'auditeur est capable d'interpréter. En particulier, l'auditeur perçoit la force de voix du locuteur, ce qui influence sa perception de la distance.

2.4 Conclusion

Un locuteur est capable de contrôler finement ses signaux vocaux, non seulement pour transmettre des informations linguistiques, mais surtout pour agir sur sa relation à l'autre : il peut aussi bien « frapper » ses interlocuteurs, que les « caresser ». Par ailleurs, comme dans le cas d'une interaction tactile, une interaction vocale n'est pas à sens unique : lorsqu'il « touche », le locuteur est touché en retour par son interlocuteur, puisque sa voix révèle, en plus de ce qu'il veut dire, une somme considérable d'informations le concernant : son identité, son groupe socio-culturelle, ses socio-affects, sa position dans l'espace… C'est en cela que l'on peut parler de « toucher vocal ».

À travers les études présentées, il apparaît cependant une limite méthodologique, que Francis Rumsey résume comme ceci :

> *Dans les expériences psychoacoustiques, il y a presque toujours une tension entre la validité écologique et le contrôle scientifique des variables – plus les variables expérimentales sont contrôlées pour observer des effets individuels, moins l'expérience est valide du point de vue écologique. Il y a ainsi une forme de principe d'incertitude à l'œuvre, puisqu'il est possible d'obtenir soit un résultat non ambigu avec une certitude élevée, mais une faible validité écologique, soit un résultat plus incertain, avec une haute validité écologique. Plus une expérience se rapproche d'une situation réelle, moins il est facile de contrôler toutes les variables.*[15] *(Rumsey 2002)*

Ainsi, il peut être difficile d'extrapoler les résultats d'expériences obtenues en conditions de laboratoire pour comprendre ce qui se produit dans la « vraie vie de tous les jours ». Or, cette compréhension est vitale dans le cas de technologies sensées permettre à des personnes isolées de conserver leurs liens sociaux à distance. Dans le chapitre suivant, nous présenterons la méthode que nous avons choisie pour tenter de concilier rigueur scientifique et validité écologique des résultats.

[15] « In psychoacoustic experiments there is nearly always a tension between ecological validity and scientific control of variables—the more tightly one controls experimental variables in order to observe individual effects, the less ecologically valid the experiment becomes. There appears to be a form of uncertainty principle at work, in that one can obtain an unambiguous result with high certainty but low ecological validity, or a more uncertain result with higher ecological validity. The more like a real-world situation the experiment becomes, the less easy it is to control all the variables. »

Chapitre 3 :
DÉMARCHE ET MÉTHODOLOGIE

Ce chapitre est consacré à la description de notre stratégie expérimentale. Nous reviendrons d'abord sur les expériences sur la localisation des sources sonores présentées précédemment, afin d'en mettre en évidence les principaux choix méthodologiques. Puis nous tracerons les grands principes de la démarche méthodologique et épistémologique qui a guidé nos expériences. Enfin, nous montrerons l'intérêt d'utiliser un robot de téléprésence pour étudier le toucher social.

3.1 Étude de la parole en milieu protégé

À travers l'exemple de la psychoacoustique, nous allons revenir sur deux points méthodologiques qui concernent toutes les études sur l'humain : le choix de la tâche, et le choix du lieu d'expérimentation. Nous examinerons la manière dont ces choix influencent la pertinence des résultats obtenus.

3.1.1 Définition de la tâche

Une problématique importante en psychoacoustique concerne le choix de la tâche demandée aux sujets pour évaluer leurs performances en localisation. En effet, la mesure ne peut être qu'indirecte, puisqu'il est impossible d'accéder directement à l'« image » acoustique qu'un auditeur se fait de son environnement (Wightman 1990).

(Recanzone et al. 1998) identifient deux paradigmes d'expérimentation possibles pour les études sur la perception. L'un consiste à demander au participant de percevoir un changement entre deux stimuli. Par exemple, une série de stimuli est diffusée via des haut-parleurs, et le participant doit indiquer tout changement de direction. Il s'agit alors d'une perception **relative**. Au contraire, on peut demander au sujet de juger les stimuli un par un : à chaque fois qu'il entend un son, il doit indiquer sa direction. Dans ce cas, il s'agit d'une perception **absolue**.

Une fois ce paradigme choisi, encore faut-il décider de la manière dont le sujet indique ses réponses. À nouveau, il existe deux possibilités : soit le nombre de choix disponible est **discret**, soit il est **continu**. Dans le premier cas, le participant n'a le choix qu'entre un nombre fini de réponses, qu'il connaît à l'avance. Il existe alors plusieurs manières de recueillir ses réponses : le sujet peut par exemple appuyer sur un bouton, tirer un levier, parler, ou même écrire. Pour évaluer ses performances, il suffit de comparer ses réponses à la réalité : l'évaluation est uniquement binaire ; soit le sujet a raison, soit il a tort. Dans certains cas, il n'est pas possible ou désirable de se limiter à un ensemble fini de réponses. D'autres solutions doivent donc être trouvées pour permettre au sujet de répondre. Pour la localisation en azimut, une des méthodes les plus courantes consiste à demander aux sujets de tourner la tête vers la source sonore, ou d'utiliser un pointeur, dans le cas où ils doivent rester immobiles. Ces relevés continus possèdent une marge d'erreur, qui dépend du matériel utilisé pour pointer la direction.

Cette problématique du choix de la méthode de réponse est particulièrement importante dans le cas des études sur la localisation absolue de la distance. En effet, la plupart des gens ne savent pas nommer précisément les distances physiques, ce qui ne signifie pas qu'ils ne sont pas capables de faire la distinction entre une source sonore située à 5 ou 8 m par exemple. Ainsi, des marqueurs visuels sont parfois mis en place pour les aider. Or, ces marqueurs peuvent influencer les réponses des sujets (Calcagno et al. 2012). Plus rarement, la distance est indiquée à l'aide d'un pointeur, comme dans le cas de (Mershon 1997). Plus originaux, (Loomis et al. 1998) ont utilisé une méthode issue des expériences en perception visuelle : il s'agit de demander aux sujets de marcher jusqu'à la position de la source sonore. Ils constatent que cette méthode fournit des résultats comparables à ceux obtenus lorsque les réponses sont verbales, mais avec une variation inter-sujets plus faible.

Il est donc important de bien choisir la tâche, non seulement en fonction de ce que l'on souhaite démontrer, mais également en fonction de ce que les sujets sont capables de réaliser.

3.1.2 Choix du lieu d'expérimentation

Par ailleurs, il existe différents lieux d'expérimentations, dont nous allons décrire les principales caractéristiques.

Tout d'abord, l'expérimentation peut se dérouler dans une pièce ordinaire, dite **réverbérante**. Une telle pièce n'est jamais parfaitement silencieuse : on peut entendre en permanence un bruit de fond, provenant de sources sonores situées à l'extérieur de la pièce, qui n'ont pas été entièrement atténuées par les murs ou les fenêtres. En outre, la propagation des ondes sonores à l'intérieur de la pièce est très complexe à modéliser, puisqu'elles sont en partie absorbées et réfléchies par chaque surface. Ces échos modifient le timbre de la voix du locuteur, et donc sa perception autophonique et sa production de parole : si l'environnement est particulièrement réverbérant, comme dans le cas d'une église par exemple, le locuteur a tendance à parler plus lentement, probablement pour que les échos aient le temps de s'atténuer entre chaque syllabe (Pelegrín-García et al. 2011). Une grandeur simple pour caractériser un tel environnement est le temps de réverbération : il s'agit de la durée nécessaire pour que le niveau d'intensité acoustique diminue de 60 dB après une impulsion sonore. Ce temps de réverbération dépend de la taille et de la forme de la pièce, mais peut-être également manipulé en modifiant son ameublement pour limiter le nombre de surfaces réfléchissantes. Par exemple, le temps de réverbération est en général plus important dans une salle de bain carrelée, que dans salon moquetté. On peut parler de salle semi-réverbérante, dans le cas des matériaux absorbants sont disposés uniquement sur un côté de la pièce, pour limiter les réverbérations.

Une **chambre anéchoïque**, ou une chambre sourde, est une salle conçue pour limiter au maximum la présence d'échos. Elle est isolée des bruits extérieurs, et ses parois sont recouvertes de pyramides ou de polyèdres en matériau absorbant, qui empêche les ondes sonores de se réfléchir. Elle permet ainsi de reproduire les conditions acoustiques théoriques dites en « **champ libre** ». Ces chambres sont réputées pour mettre mal à l'aise les personnes qui s'y trouvent, puisque le silence y est tel qu'elles perçoivent des bruits de leur corps habituellement inaudibles.

Une autre manière d'approcher ces conditions en champ libre est de travailler en extérieur, loin de toute surface réfléchissante, afin que le son se propage de la même façon dans toutes les directions : par exemple, au milieu d'un champ d'herbes. Ce sont les conditions dites « **en champ ouvert** ».

L'expérimentation peut également se dérouler au casque. Dans ce cas, il existe deux familles de méthodes : celles qui reposent sur l'utilisation d'enregistrements obtenus à partir de têtes artificielles, et celles qui reposent sur des simulations à base de HRTF (conditions VAD : Virtual Auditory Display).

3.1.3 Avantages et limites des études en champ libre

Les études en champ libre sont plus simples d'un point de vue acoustique. En particulier les chambres anéchoïques permettent de contrôler parfaitement ce que l'auditeur entend, sans aléas extérieurs (ex : bruit de vent, de voitures...). Cette simplicité peut se révéler comme un avantage, ou comme un inconvénient, en fonction de ce que l'on cherche à étudier.

Les premières études sur la localisation spatiale concernaient la perception de l'azimut (et de l'élévation dans une moindre mesure). Elles se sont faites essentiellement en chambre anéchoïque, ou en champ ouvert, à l'aide de plusieurs haut-parleurs identiques disposés en arc de cercle autour du sujet. Dans leur revue, (Middlebrooks, Green 1991) choisissent tout simplement d'écarter les expériences se déroulant dans d'autres conditions acoustiques (au casque, ou en présence de réverbération).

En effet, l'intérêt d'utiliser des haut-parleurs plutôt qu'un casque est qu'il n'y a pas d'ambiguïté dans la question posée au sujet : le son provient véritablement de sources différentes. Ainsi, il suffit de comparer la direction indiquée par le sujet à la position réelle du haut-parleur. Lorsque l'écoute se fait au casque en revanche, la direction du son perçue par les sujets est toujours artificielle : ce sont les stimuli eux-mêmes qui portent l'information spatiale. On peut donc toujours se demander si les résultats obtenus sont réellement représentatifs de l'audition humaine, ou sont au contraire biaisés par le dispositif de mesure. Ainsi, (Møller et al. 1999) et (Minnaar et al. 2001) ont pu mettre en évidence des performances en cas d'utilisation de têtes artificielles inférieures à celles obtenues avec des enregistrements faits avec des têtes humaines, même différentes de l'auditeur. En cas d'utilisation de HRTF spécifiques à l'auditeur, il est possible d'obtenir des stimuli binauraux quasiment indiscernables de ceux produits par de véritables sources sonores (Langendijk, Bronkhorst 2000). Cependant, mesurer ces HRTF pour chaque sujet dans le cadre de tests psychoacoustiques reste une étape très contraignante. Les défis actuels dans ce domaine de recherche concernent la personnalisation des HRTF à partir d'une base de données préenregistrées, et l'interpolation en temps-réel, afin de pouvoir tenir compte des mouvements de tête de l'auditeur.

Du côté des études sur la perception de la distance, le problème s'est posé différemment. En effet, tant que les tests imitaient ceux de localisation en azimut (c'est-à-dire en champ libre, avec des signaux artificiels), le seul indice disponible pour les sujets était l'intensité perçue. Cette intensité pouvait être modifiée artificiellement, sans pour autant changer la distance de la source sonore. Il était donc possible de faire des tests avec un seul haut-parleur, ce qui réduit

considérablement le matériel de tests. Cependant, un tel dispositif pouvait entrer en conflit avec d'autres processus cognitifs des auditeurs, en particulier dans le cas où l'expérience se produisait dans une petite salle et que les distances simulées étaient très grandes. Ainsi, les sujets étaient parfois incités à rapporter la distance « apparente » de la source sonore. C'est le cas notamment dans les études de (Gardner 1969) et (Philbeck, Mershon 2002). Cette consigne est souvent évoquée dans les études sur la perception de la distance, en faisant référence à la revue de (Carlson 1977), concernant les instructions à utiliser dans les études sur la perception visuelle : avec la précision distance « apparente », les sujets s'appuieraient plus sur les facteurs perceptuels que sur les facteurs cognitifs. Cette consigne est très étrange, puisqu'elle demande aux sujets de mettre en doute leur propre perception, et de donner des réponses parfois absurdes : par exemple, dire que la source du son semble se trouver à 10 m, alors que la profondeur de la pièce n'est que de 5 m. Cela pourrait avoir biaisé les résultats en exagérant l'importance de la force de voix ; les sujets comprenant implicitement que les voix chuchotées doivent être perçues plus proches que les voix modales, et les voix criées plus éloignées.

Par ailleurs, en modifiant l'intensité d'un enregistrement, on crée une voix schizophonique[16], qui n'aurait pas pu être produite par un être humain, incapable de faire varier l'intensité de sa voix sans en modifier dans le même temps ses autres caractéristiques. Si les voix chuchotées sont perçues plus proches, c'est peut-être tout simplement parce que si elles avaient été produites par un être humain, et non par un haut-parleur, alors elles ne seraient pas audibles à grande distance. À l'inverse, un locuteur n'a pas besoin de crier pour s'adresser à une personne proche de lui. Comprendre qu'un cri, même de forte intensité, vient d'une distance proche, est donc contre-intuitif.

Ces résultats sont donc valides et intéressants, car ils mettent en évidence le fait que les auditeurs utilisent leur expérience a priori pour déterminer la distance des sources sonores. Cependant, ils ne nous renseignent pas beaucoup sur la manière dont un auditeur perçoit la distance des sources sonores dans sa vie quotidienne. Pour progresser dans la compréhension de ces mécanismes, il a fallu tenir compte de l'environnement acoustique, en plaçant les expériences dans des pièces ordinaires ou virtuelles. C'est grâce à ces expériences que l'importance des indices acoustiques liés à l'environnement, tel que le ratio d'énergie directe/réverbérée, a pu être mise en évidence.

3.1.4 Résumé

A priori, les études en chambre anéchoïque peuvent paraître plus rigoureuses sur le plan scientifique, puisqu'elles sont hautement reproductibles : aucun son ne peut entrer dans une chambre anéchoïque sans que les expérimentateurs le sachent. Cependant, cet environnement est si artificiel que certains phénomènes ne peuvent pas s'y produire. Le critère de reproductibilité ne doit donc pas être le seul critère permettant de juger la validité d'une expérience : il faut également s'assurer que la tâche demandée aux sujets et le contexte dans

[16] Expression empruntée à (Schafer 1977), qu'il utilise pour désigner la séparation entre un son original et sa reproduction électroacoustique, qui échappe au contexte dans lequel le son a été produit pour le placer dans un autre lieu et un autre moment.

lesquels ils sont placés se rapprochent suffisamment de leur réalité écologique, c'est-à-dire de leurs conditions de vie réelle.

3.2 Pour une écologie des usages

Notre objectif est de parvenir à étudier le toucher vocal de la manière la plus écologique possible en conditions de laboratoire. Dans cette partie, nous allons présenter les spécificités de la méthodologie utilisée au cours de cette thèse. Nous verrons que cette méthodologie repose sur trois piliers : un lieu d'expérimentation inspiré des Living Lab, un scénario expérimental sophistiqué, et un cadre de pensée holistique.

3.2.1 Le Living Lab Domus

Notre méthodologie s'appuie d'abord sur un lieu particulier : le Living Lab Domus. Un Living Lab est un lieu de rencontre entre les différents acteurs de la société (3P : Public, Private, People). Il est conçu pour permettre le développement de technologies innovantes et éthiques, dans un processus de co-construction agile entre les fabricants de technologies, et les utilisateurs : plutôt que de suivre toutes les étapes de développement d'un prototype, avant de le faire tester, l'idée est de faire du développement incrémental, en soumettant régulièrement des propositions aux utilisateurs, qui vont pouvoir orienter dès le début le prototypage. Un exemple de co-construction menée pendant cette thèse sera développé au chapitre 6.

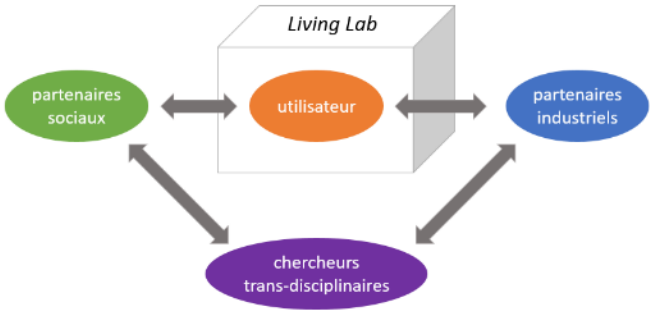


Figure 53 : Le Living Lab : un lieu de rencontre pour la recherche et l'innovation

Domus est une plateforme d'expérimentation du laboratoire d'informatique de Grenoble. Situé initialement dans le Centre des Technologies du Logiciel, il a été déplacé en 2019 dans la Maison de la Création et de l'Innovation. Les plans du bâtiment correspondant sont visibles en Figure 54 et Figure 55. Outre quelques bureaux et lieux de stockage, Domus contient plusieurs salles dédiées à l'expérimentation. La chambre sourde, et une salle isolée phonétiquement, qui permet des enregistrements faiblement bruités. La plateforme expérimentale est une salle modulable, équipée de micros et de caméras, qui peut être réaménagée en fonction des besoins des expérimentateurs. L'appartement domotique se rapproche d'un véritable appartement, tout en étant équipé d'un système de capteurs et d'actionneurs. Enfin, la régie permet de coordonner les enregistrements issus des salles d'expérimentations.

Cette plateforme est donc construite spécifiquement pour l'expérimentation écologique : il s'agit de pouvoir tester les technologies dans des lieux similaires à ceux dans lesquels elles vont être utilisées ; tout en disposant d'un ensemble de capteurs permettant de recueillir les données expérimentales.

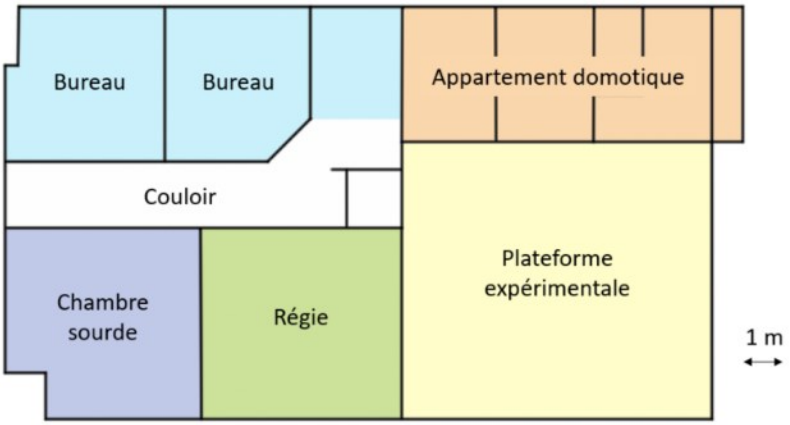


Figure 54 : Plan de l'ancien Domus, dans le bâtiment CTL (Centre des Technologies du Logiciel)

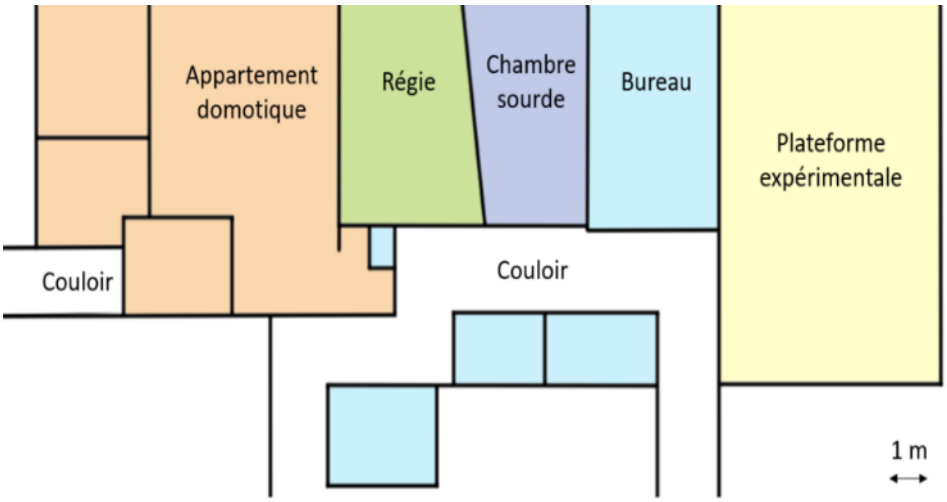


Figure 55 : Plan du nouveau Domus, dans la MaCI (Maison de la Création et de l'Innovation)

La plateforme Domus entretient une collaboration étroite avec le FabMSTIC, le fablab universitaire, financé par le LIG (cf. Figure 56). C'est un lieu de prototypage, ouvert à tous les étudiants, enseignants et chercheurs de l'université (y compris non informaticiens), ainsi qu'aux entreprises privées sous contrat avec l'université. Il est constitué d'un open space, où les utilisateurs du fablab peuvent venir travailler, et de quatre ateliers. L'atelier électronique contient le nécessaire pour pouvoir réaliser un système électronique : composants (y compris circuits imprimés), voltmètres, pinces, fers à souder... L'atelier vinyle et impressions 3D permet de découper des formes dans du papier vinyle (tel que des logos), ou d'imprimer des objets en trois dimensions à partir de modèles numériques. Le troisième atelier contient une

thermoformeuse, qui permet par un procédé de chauffage de donner à une plaque de plastique la forme d'un moule ; ainsi qu'une découpeuse laser, qui permet de découper ou graver des plaques de plastique ou de bois aggloméré de 1 à 3 cm d'épaisseur, suivant un tracé numérique. Enfin, l'atelier mécanique contient des outils de bricolage plus classiques (visserie, marteaux, colles, scie…).

Figure 56 : Photo prise au FabMSTIC (Université Grenoble Alpes)

C'est au sein de ce fablab qu'a été conçu le robot de téléprésence Robair. Développé depuis 2012, il s'enrichit d'année en année, au grès de différents projets étudiants et projets de recherche. Trois versions de ce robot sont visibles en Figure 57. Ses plans de construction, ainsi que son programme informatique, sont disponibles en ligne sur le dépôt Git du fabMSTIC (https://github.com/fabMSTICLig/RobAIR).

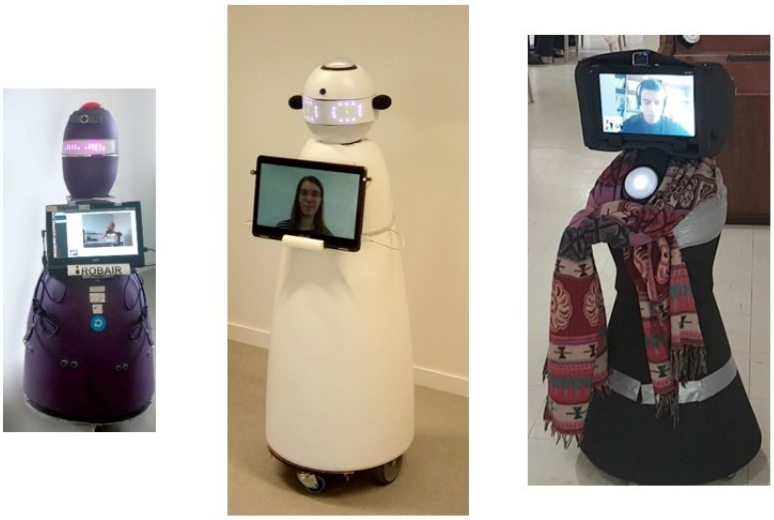


Figure 57 : Trois versions du robot Robair
À gauche : première version, photographiée en 2015
Au milieu : notre version du robot
À droite : version développée par les étudiants du master Manintec (Management de l'Innovation et des Technologies) dans le cadre du projet #FromLivingLab en collaboration avec des résidents du foyer pour personnes âgées Roger-Meffreys

3.2.2 Des scénarios pour une recherche écologique

Par ailleurs, notre méthodologie se base sur la construction de scénarios. Comme dans le cas d'expériences en psychologie sociale, il s'agit de faire en sorte que les sujets ne se doutent pas de ce que nous cherchons à étudier. Dans le cas contraire, leur intelligence pourrait biaiser les résultats expérimentaux. En effet, ils pourraient exagérer leurs comportements dans le sens qu'ils imaginent attendu par les chercheurs, pour le faire plaisir, ou par désir de respecter des consignes implicites. Au contraire, s'ils sont plutôt réticents face à l'expérimentation, ils pourraient inhiber leur comportement : par exemple, en s'efforçant d'avoir un visage impassible s'ils savent qu'on cherche à étudier leurs émotions faciales. Dans les deux cas, ils ne se contentent pas de réaliser la tâche qui leur est demandée, mais s'observent en train de réaliser la tâche. Ils sont dans un état de **métacognition** : les processus cognitifs mis en jeu sont donc probablement différents, de ceux qui se produiraient hors contexte expérimental.

En psychologie, cet effet est connu sous le nom d'**effet Hawthorne**. On parle également de **paradoxe de l'observateur**, mis en évidence en socio-linguistique par Labov (1970) : puisque les gens modifient leur comportement lorsqu'ils se savent observés, il faudrait pouvoir observer leur comportement quand ils ne sont pas observés. En pratique, on essaiera de limiter cet effet d'observation, en plaçant les locuteurs dans un contexte familier, et en dissimulant les buts de l'expérience, afin qu'ils ne puissent pas se douter de ce qu'on est en train d'observer.

Nous attachons un soin particulier à cette mise en scène : il ne s'agit pas simplement de faire en sorte que les sujets ignorent ce qu'ils sont en train de faire ; leur attention doit être entièrement tournée vers une tâche prétexte, afin que les performances qu'ils réalisent soient les plus proches possibles de celles qu'ils réaliseraient dans la vie de tous les jours. Autrement dit, notre objectif est d'observer nos sujets se comporter de manière naturelle dans des circonstances exceptionnelles, qui sont celles d'un laboratoire de recherche.

3.2.3 Un cadre holistique

Nous allons à présent évoquer le paradigme épistémologique dans lequel s'inscrit notre thèse, à travers quelques références historiques.

Depuis les travaux de Norbert Wiener, fondateur de la cybernétique, il est admis que la communication met en jeu des boucles de **rétroactions** (Wiener 1948). L'auditeur n'est donc pas un élément passif de la communication : il renvoie en permanence des informations au locuteur qui intègre ce *feedback* dans son comportement. Ce modèle extrêmement général est à la base de l'automatique et permet de modéliser n'importe quel système de contrôle et de régulation.

Wiener a notamment inspiré l'école de Palo Alto, née dans les années 50, qui propose une vision « orchestrale » de la communication, pour reprendre l'expression de (Winkin 1981). Ainsi, la communication n'est plus perçue comme l'affaire d'individus isolés les uns des autres s'envoyant des messages à l'aveugle : tous font partie d'un système, tels les musiciens d'un orchestre qui s'écoutent les uns les autres pour jouer juste. Ils ne se contentent pas de communiquer, mais plutôt **participent à la communication** (Watzlawick et al. 1972).

A la même époque, l'anthropologue Edward T Hall enseigne des techniques de communication interculturelle aux employés du ministère des affaires étrangères des Etats-Unis. Quelques années plus tard, il développera sa théorie **proxémique**, portant sur la manière dont les êtres humains occupent et appréhendent l'espace. Lui aussi s'intéresse à la notion de système d'interactions, comme le montre cette citation traduite en français :

> *Les découvertes des éthologistes et des psychologues animaliers suggèrent que : a) chaque organisme habite son propre monde subjectif, qui est une fonction de ses perceptions sensorielles, et la séparation arbitraire de l'organisme de ce monde modifie le contexte et en déforme ainsi le sens et b) la ligne de démarcation entre l'environnement intérieur et extérieur de l'organisme ne peut pas être identifiée précisément. La relation organisme-biotope ne peut être comprise que si elle est vu comme une collection délicatement équilibrée de mécanismes cybernétiques dont les rétroactions positives et négatives exercent un subtile mais continu contrôle sur la vie.* ***Autrement dit, l'organisme et son biotope constitue un système unique et cohérent*** *(compris dans une collection de systèmes plus vastes). Étudier l'un sans référence à l'autre est dénué de sens. (Hall et al. 1968)*[17]

Le système d'interaction ne se limite donc pas simplement à des individus communiquant, mais englobe également une dimension culturelle, sociale et environnementale. La parole a donc un **contexte**. Les réflexions de Hall suggèrent également que la parole ne se limite pas au domaine acoustique, mais intègre d'autres sens : la vue, le toucher et même l'odorat. Elle est **multisensorielle**.

Dans les modèles modernes, la parole ne se limite donc plus à un simple signal acoustique, mais englobe un ensemble de contextes et d'informations paralinguistiques. On peut citer notamment le modèle proposé par Liénard en 1977 (cf. Figure 58). Soulignons que dans ce modèle, les interlocuteurs partagent le même espace spatio-temporel : ils sont présents au même endroit, et au même moment. En outre, il ne fait pas de la communication une fin en soi : A parle à B pour susciter une action de la part de B. Il y a donc un **enjeu** à la communication qui va au-delà du simple échange de messages.

La communication parlée est donc conçue non pas comme un match de ping pong entre deux individus, où la balle serait le message linguistique, mais comme une interaction multimodale entre les locuteurs et leur environnement. Si au cours de cette thèse, nos études ne porteront que

[17] « The findings of ethologists and animal psychologists suggest that : (a) each organism inhabits its own subjective world, which is a function of its perceptual apparatus, and the arbitrary separation of the organism from that world alters context and in so doing distorts meaning ; and (b) the dividing line between th organism's internal and external environment cannot be pinpointed precisely. The organism-biotope relationship can only be understood if it is seen as a delicately balanced series of cybernetic mechanisms in which positive and negative feedback exert subtle but continuous control over life. *That is, the organism and its biotope constiute a single, cohesive system* (within a series of larger systems). To consider one without reference to the other is meaningless. »

sur les signaux vocaux, ce sera pour montrer en quoi ceux-ci sont porteurs non seulement d'informations acoustiques, mais aussi spatiales.

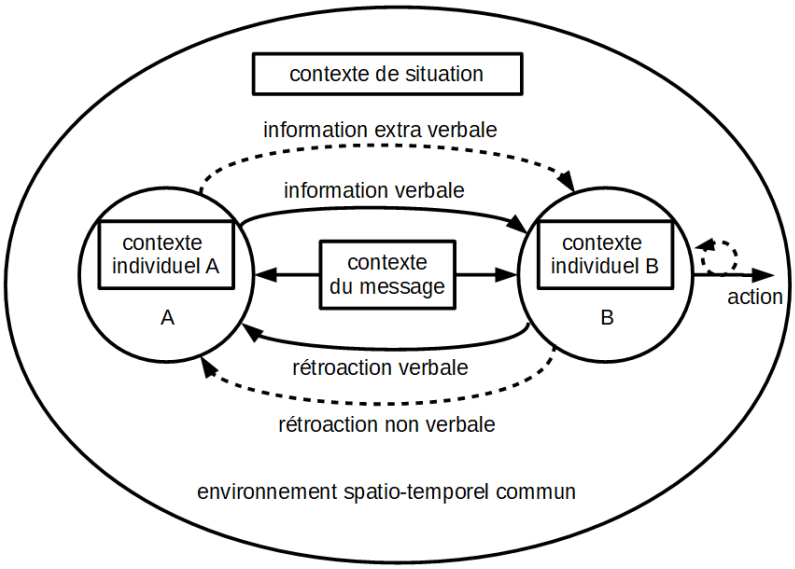


Figure 58 : Schématisation de la communication parlée, d'après (Liénard 1977)

3.2.4 Résumé

Nous venons de présenter les trois piliers sur lesquels s'appuie notre méthodologie. Il s'agit tout d'abord d'un lieu d'expérimentation spécifique : le Living Lab Domus, en étroite collaboration avec le fablab universitaire, FabMSTIC. Il s'agit également d'une manière de concevoir les expériences, dans le but d'observer les comportements les plus écologiques possibles. Enfin, cette démarche expérimentale prend sa source dans une conception holistique des interactions humaines : le locuteur n'est pas vu comme un individu isolé des autres et de son environnement.

3.3 Le robot de téléprésence comme instrument de recherche

Pour finir ce chapitre sur la méthodologie, nous allons expliquer l'intérêt d'utiliser des robots de téléprésence pour étudier les interactions humaines, en dépit des artefacts technologiques produits par ses robots.

3.3.1 Contexte scientifique

Un robot de téléprésence parfait permettrait à son utilisateur de percevoir et agir dans l'environnement local comme s'il y était présent. Comme l'apparence du robot a un impact sur l'interaction, il serait nécessairement un clone de son utilisateur, de sorte que ses interlocuteurs ne seraient pas capables de faire la différence entre le robot de téléprésence et son propriétaire. Dans ces conditions, le pilote du robot n'aurait aucune difficulté à contrôler son toucher vocal à distance ; et notre problématique de thèse ne se poserait plus.

Or, de tels robots n'existent pas encore ; et n'existeront probablement jamais. Dans cette thèse, ce sont les robots de téléprésence actuels qui nous intéressent : des robots aux capacités limitées, et qui placent leurs utilisateurs dans une situation d'ubiquïté. Ces robots induisent des interactions particulières, puisque les interlocuteurs ne se comportent pas en téléprésence, comme ils le feraient en présentiel. Étudier ces interactions a donc d'abord un intérêt industriel, puisque cela devrait permettre de trouver des pistes d'amélioration. Il y a également un intérêt scientifique, puisque le robot permet à la fois d'instrumenter l'interaction, et d'en révéler les mécanismes.

3.3.2 Instrumentation de l'interaction

Un robot de téléprésence embarque plusieurs capteurs (au minimum une caméra et un microphone). Il est donc possible de conserver des traces de l'interaction, sans avoir besoin d'équiper les utilisateurs de capteurs supplémentaires. L'interaction est également simplifiée, puisqu'il est possible de savoir exactement ce que le pilote perçoit et produit à chaque instant, car tous ses signaux passent par le robot. En outre, l'expérimentateur n'a pas à choisir de point de vue pour les enregistrements audio ou vidéo : les traces recueillies sont parfaitement fidèles à l'expérience des utilisateurs. Le robot de téléprésence constitue donc une véritable plateforme d'expérimentation mobile, qu'il est possible de sortir du laboratoire pour étudier les interactions dans leur milieu naturel (ex : Ehpad, bureaux ou école).

3.3.3 Révélateur des mécanismes de l'interaction

Par ailleurs, du fait de ses limitations, le robot de téléprésence permet de mettre en évidence les mécanismes de l'interaction. Ainsi, en comparant une interaction en téléprésence et une interaction en présentiel, nous cherchons à comprendre quels sont les signaux physiques vecteurs de l'interaction. Par analogie, c'est en comparant l'anatomie d'un oiseau qui ne vole pas à celle d'un oiseau qui vole que l'on peut saisir ce qui permet à l'oiseau de voler. En mettant en défaut la communication, le robot de téléprésence permet donc d'agir comme un révélateur.

Ce révélateur est d'autant plus nécessaire que l'expérimentateur est déjà un expert en interaction, au même titre que ses sujets : il interagit au quotidien, depuis sa naissance. Ses interactions sociales sont même nécessaires à sa survie et à son développement, comme en témoignent les cas de jeunes enfants isolés de leur famille, étudiés notamment par (Ainsworth, Bowlby 1991), auteurs de la théorie de l'attachement, ou encore les expériences atroces de Harry Harlow sur les jeunes macaques séparés de leur mère, qui, bien que correctement nourris, finissent par dépérir en l'absence de toucher maternelle (Ovadia 2016). Le regard du chercheur en interaction ne peut donc en aucun cas être extérieur à son objet d'étude, mais est au contraire chargé d'a priori et d'intelligence naïve dont il n'a pas forcément conscience.

3.3.4 Résumé

Le robot de téléprésence est donc un instrument de recherche : c'est un outil de télécommunication accessible au grand public, conçu sur la base de connaissances scientifiques pluridisciplinaires ; et qui permet de développer cette base de connaissance (Mandran 2017).

3.4 Conclusion

Ce chapitre nous a permis de présenter de façon très générale la démarche méthodologique mise en œuvre au cours de cette thèse. Son objectif est de pouvoir étudier l'humain dans des conditions **écologiques**, c'est-à-dire les plus proches possibles de leur milieu d'interaction naturel. Elle repose sur la conception d'un **scénario expérimental**, destiné à détourner l'attention des sujets pour les empêcher d'avoir une métacognition de leur boucle perception-action. Nous nous sommes également appuyés sur un lieu dédié à l'expérimentation et à l'innovation pluridisciplinaires : le **Living Lab Domus**, associé au **fablab universitaire FabMSTIC**. Le robot de téléprésence Robair, développé au sein de ces deux structures, constitue un **instrument de recherche** pour étudier le toucher social.

Nous reviendrons sur des exemples plus concrets de mise en pratique de cette méthodologie dans les chapitres 4 à 6, consacrés à nos expériences.

Chapitre 4 :
DISTANCE SOCIALE, PORTÉE VOCALE ET PERCEPTION DE L'ESPACE

Dans le chapitre 2, nous avons développé la notion de **toucher vocal**, en proposant un aperçu de la vaste étendue des productions vocales accessibles à un locuteur. Ainsi, une personne qui parle échange une immense quantité d'informations avec ses interlocuteurs, à la fois linguistiques et paralinguistiques, tout en contrôlant en permanence sa **portée vocale** pour s'adapter aux contraintes de l'environnement acoustique (distance et bruit). Cette portée vocale participe grandement au toucher social, puisqu'elle est fondamentalement liée à la manière dont les interlocuteurs occupent socialement et culturellement l'espace (**proxémie**) : en particulier, pour pouvoir s'entendre, la portée vocale doit être cohérente avec la distance interpersonnelle. Elle dépend également fortement du **contexte social et culturel** : par exemple, deux personnes assises côte à côte n'utilisent pas la même portée vocale si elles se trouvent à la bibliothèque, ou dans un bar (même silencieux).

Ce chapitre est dédié à la question suivante : nous nous demandons si des **variations socio-affectives** sont susceptibles, indépendamment des variations de **portée vocale** qu'elles engendrent, d'influencer la manière dont un auditeur perçoit la posture du locuteur, et en particulier sa **distance physique**. En effet, à travers la perception des socio-affects, l'auditeur pourrait percevoir une **distance sociale**. Si cette distance sociale n'est pas forcément en relation avec la distance physique qui sépare les interlocuteurs, on constate néanmoins que certains socio-affects apparaissent de manière privilégiée sous certaines conditions de distance physique. Par exemple, pour pouvoir réconforter quelqu'un, il est nécessaire d'être suffisamment proche pour pouvoir le toucher et lui parler d'une voix douce. Ainsi, une voix douce et réconfortante est intrinsèquement associée à une courte distance. Au contraire, la voix autoritaire d'un enseignant serait plutôt associée à une grande distance, puisqu'elle s'adresse toute une classe.

Pour pouvoir étudier cette question de manière méthodique, et tout en sachant que l'effet recherché est probablement très faible, nous avons conçu plusieurs nouvelles expériences. Elles consistent à demander à un auditeur de repérer acoustiquement où se trouve son interlocuteur dans l'espace. Cet interlocuteur est en réalité une **locutrice expérimentée** à la création de corpus pour l'étude de la prosodie socio-affective. Au cours de l'expérience, elle se déplace dans la pièce, et modifie sa manière de parler pour produire deux socio-affects opposés. Ces socio-affects étant intrinsèquement associés à deux intensités différentes, la locutrice contrôle également sa portée vocale pour produire deux forces de voix différentes. Nous chercherons à quantifier l'influence de ces **quatre types de voix** (combinaison d'un socio-affect et d'une intensité) sur la perception des participants.

L'étude se compose de trois expériences. Pour les deux premières, le sujet est présent dans la salle de test, et interagit directement avec la locutrice. Dans un cas, une mise en scène complexe

est utilisée pour détourner l'attention des sujets, et les empêcher de deviner que la locutrice modifie sciemment sa manière de parler. Dans l'autre cas au contraire, on demande aux sujets de prêter attention à ces variations. Enfin, la troisième expérience se déroule en téléprésence : le sujet écoute à distance la locutrice, dont la voix a été enregistrée via un robot de téléprésence.

4.1 Premier test psychoacoustique

Une première expérience a permis de poser le cadre de notre étude sur la perception de la portée vocale. Le principe était de demander à un sujet aux yeux bandés de localiser dans l'espace une personne prononçant des mots isolés. Pour cette expérience, nous avons choisi de ne pas utiliser de haut-parleurs comme sources sonores, et de mettre en place un scénario sophistiqué, afin que le sujet pense véritablement interagir avec une personne, et ne puisse pas se douter des buts réels de l'expérience. Dans cette partie, nous présenterons en détail notre méthodologie, ainsi que la mise en scène choisie pour détourner l'attention des sujets ; puis nous passerons à l'analyse des résultats.

4.1.1 Méthodologie

L'objectif principal de l'expérience était d'étudier l'influence de deux facteurs vocaux (intensité et socio-affect), sur la perception de trois variables spatiales (direction, distance et orientation). Ces dernières correspondent aux informations principales permettant de décrire la position relative d'une personne qui me parle : De quel côté se trouve-t-elle ? À quelle distance ? Est-elle tournée vers moi ? Les valeurs prises par ces différents facteurs et variables sont récapitulées dans le Tableau 4, et expliquées dans la suite.

Tableau 4 : Facteurs et variables étudiées

Facteurs vocaux		**Variables spatiales**		
socio-affects	**intensité**	**direction**	**distance**	**orientation**
doute confiance	faible forte	avant gauche droite derrière	proche (1m70) milieu (2m50) loin (3m30)	face dos

4.1.1.1 Choix des intensités et socio-affects

Les socio-affects ont été sélectionnés par Aubergé, sur la base de ses précédents travaux en prosodie audio-visuelle. Citons en particulier (Shochi et al. 2009), qui présente une étude sur la perception interculturelle d'attitudes produites par des locuteurs natifs anglais, japonais et français. Les douze attitudes du corpus français, produit par Aubergé, ont été validées perceptivement par 30 sujets français.

Les deux socio-affects choisis se caractérisent par une opposition : l'un est un geste doux, qui doit rapprocher l'auditeur, l'autre est un geste dur, qui doit le repousser. Initialement, le choix s'est porté sur le couple politesse vs. autorité. Cependant, ces attitudes auraient peut-être semblées artificielles pour les sujets ; ainsi en pratique, les productions d'Aubergé se

rapprochent plutôt d'un couple doute vs. confiance. Une analyse prosodique plus détaillée sera fournie en section 4.1.3.

Par ailleurs, cette opposition entre un geste doux et un geste dur se traduit par des écarts d'intensité : le premier geste étant par définition plus faible que le second. Ces variations d'intensité ayant très certainement un effet sur la perception spatiale des sujets, il était nécessaire d'ajouter une variable d'intensité, pour tenter de séparer les effets du socio-affect, de ceux de son intensité intrinsèque. Les productions vocales étudiées dans cette expérience sont donc divisées en quatre catégories : deux catégories « naturelles », le doute faible et la confiance forte, et deux catégories « artificielles », le doute fort, et la confiance faible.

4.1.1.2 Choix des distances

Le choix des distances étudiées a été effectué en tenant compte à la fois de la littérature et des dimensions du plateau expérimental auquel nous avions accès (7.1 x 8.6 m). L'objectif était d'étudier des distances suffisamment diverses, et donc intéressantes en terme de proxémie. Cependant, il ne fallait pas choisir des distances trop extrêmes, qui auraient pu être aisément discriminées par les sujets. Par exemple, un sujet pourrait probablement faire aisément la distinction entre un locuteur présent à 1m, 2m50 et 5m. À l'inverse, il ne fallait pas choisir des distances trop resserrées, qui sembleraient aux sujets impossibles à discriminer.

Pour définir les distances « proche » et « loin », nous nous sommes donc intéressés à la notion de flou perceptif : autrement dit, dans quelle mesure les réponses des sujets varient lorsqu'ils doivent estimer à plusieurs reprises la distance d'une source sonore. En pratique, cela correspond à l'écart-type des distances estimées par un sujet, pour une distance fixe. Or cette donnée est rarement indiquée dans la littérature : le plus souvent, on a accès uniquement à la moyenne de la distance perçue, en fonction de la distance réel. Si un écart-type est représenté, il s'agit donc de celui de l'ensemble des sujets ayant passé l'expérience, et pas d'un sujet isolé. Nous en avons sélectionné deux exemples particulièrement intéressant en terme de visualisation des données : la

Figure 59 est un schéma en vue de dessus de l'expérience ; très intuitive, elle fait apparaître à la fois la position des sources sonores et les estimations des sujets ; la Figure 60, plus abstraite, est une courbe de mesures, représentant la distance perçue en fonction de la distance réelle de la source sonore.

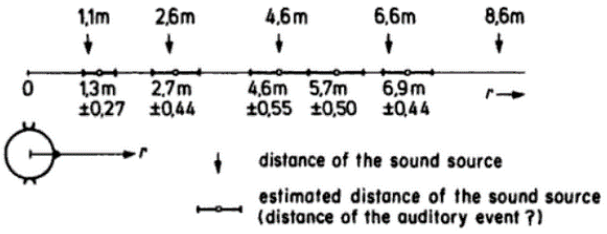


Figure 59 : Représentation schématique en vue du dessus, extraite de (Blauert 1997), d'après (Haustein 1969)

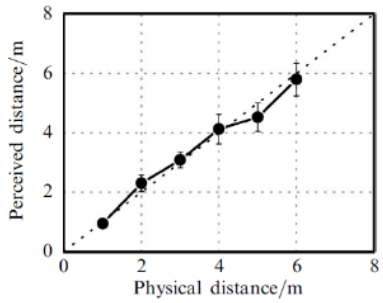


Figure 60 : Représentation sous forme de courbe (Calcagno et al. 2012)

La revue de (Kolarik et al. 2016) cite quelques estimations de la variation minimale de distance nécessaire pour percevoir un changement de distance : cette variation minimale représente entre 5 et 25 % de la distance de référence, mais varie d'une étude à l'autre. En effet, elle dépend très probablement de nombreux facteurs, tels que le type de stimuli utilisé, l'environnement de test choisi ou encore les distances étudiées.

Par ailleurs, (Zahorik et al. 2005) rapportent une estimation individuelle du flou perceptif, obtenue à partir d'une cohorte de 9 sujets dont les HRTFs ont été mesurées dans un petit auditorium : le flou perceptif irait ainsi de 20 à 60 % de la distance effective, lorsque les distances sont représentées en échelle logarithmique. On peut en déduire des estimations basses et hautes du flou perceptif pour les trois distances choisies pour notre expérience. Ces estimations sont représentées en Figure 61. Au-dessus de l'axe des distances, apparaissent également les espaces de la théorie proxémique décrite dans (Hall 1966).

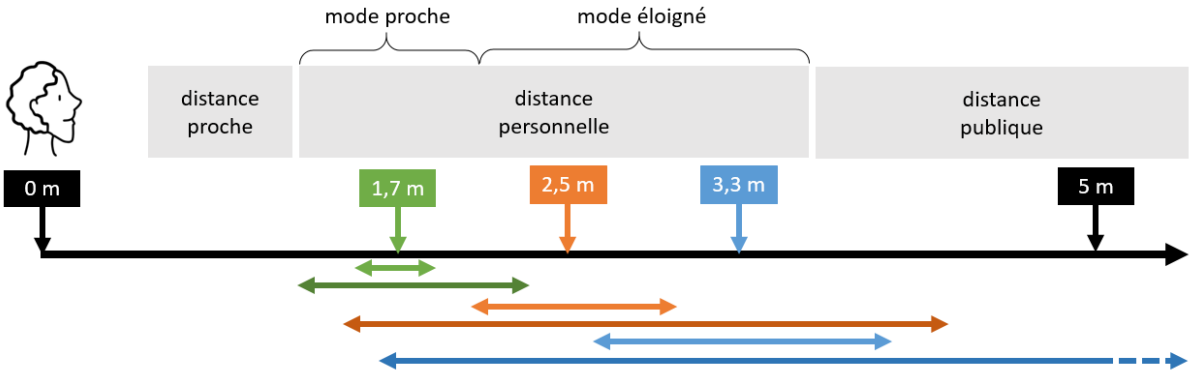


Figure 61 : Distances étudiées en regard de la théorie proxémique de Hall et des estimations basse et haute du flou perceptif à chaque distance

Par ailleurs, un phénomène intéressant se produit dans cette zone. En effet, les études en psychoacoustique ont montré que nous avons tendance à surestimer la distance des sources sonores proches, et à sous-estimer celles des sources éloignées. Or, le point de bascule se trouve probablement entre les distances étudiées. En effet, (Anderson, Zahorik, 2014) ont d'abord situé ce point à 1 m 90, puis à 3 m 22 dans une salle plus réverbérante, dans laquelle les distances sont perçues plus éloignées.

4.1.1.3 Choix des directions et orientations

Les autres variables spatiales étaient plus simples à définir. Pour la direction, nous nous sommes inspirés des axes de référence en anatomie : dorso-ventral et gauche-droite. Toute autre direction azimutale est un mélange de ces deux composantes. Pour l'orientation, la locutrice était soit tournée face au sujet, soit dos à lui.

4.1.1.4 Choix des mots-clés

La liste des mots-clés prononcés est vaste, puisqu'elle compte plus de 40 mots, de 1 à 4 syllabes. Ces mots ont été choisis en relation avec le prétexte expérimental sur lequel nous reviendrons ci-après. Ils font référence à 40 boîtes à odeurs utilisées au cours de l'expérience. Ces boîtes sont d'apparences identiques, mais issues de deux jeux de société différents[18]. En pratique, les mots étaient organisés en quatre listes de 20 mots chacune :

liste A : abricot, anis, banane, bonbon suisse, cassis, chocolat, citron, eucalyptus, lavande, loukoum, miel, menthe, orange, pamplemousse, persil, pomme, praline, rose, vanille, violette

liste B : algue, ananas, biscuit, brûlé, cacahouète, céréales, champignon, citron, citronnelle, clémentine, colle, coriandre, fraise, girofle, melon, menthe, noix de coco, orange, rose, savon

liste C : abricot, anis, banane, ~~bonbon suisse~~, cassis, chocolat, citron, eucalyptus, fleur d'oranger, lavande, ~~loukoum~~, miel, menthe, noisette, orange, pamplemousse, pin, persil, pomme, ~~praline~~, rose, vanille, violette

liste D : identique à la liste B

Les quatre listes étaient mélangées, et mises bout à bout pour constituer une liste des 64 mots-clés. Seuls les quatre derniers mots proviennent de la liste D. La liste C est identique à la liste A, à quelques mots près, de sorte que la locutrice ne répète pas exactement le même mot pour la même boîte à odeur.

Chaque boîte à odeur était étiquetée sur le dessus et sur le dessous. Les numéros du dessus correspondait à ceux de la liste A et B, et les numéros du dessous à ceux de la liste C et D. Pendant la mise en place de l'expérience, les deux séries de boîtes à odeurs étaient disposées sur la table basse de façon à faciliter leur manipulation : rangées sur quatre colonnes dans l'ordre croissant, les boîtes A/C à gauche, et les boîtes B/D à droite. Dès qu'une boîte était utilisée, elle était retournée pour la distinguer des autres.

4.1.2 Scénario de la tâche prétexte

La particularité de cette première expérience est que les sujets ne devaient pas avoir conscience de l'existence des deux facteurs vocaux : intensité et socio-affect. Or, ils auraient pu facilement s'en douter si l'expérience avait été directement présentée comme un test de localisation spatiale, ce qui leur aurait peut-être permis d'adapter leur perception en conséquence. Dans un

[18] Le Loto des Odeurs (Sentosphere) et Les Boîtes à Odeurs (Nature et Découverte)

premier temps, nous avons donc préféré utiliser une mise en scène pour détourner l'attention des sujets.

4.1.2.1 Prétexte

Les sujets étaient recrutés pour participer à une expérience sur le goût et l'odorat. Cette expérience était censée être le fruit d'une collaboration entre le LIG, le Gipsa-Lab et un laboratoire japonais, dont l'objectif était de concevoir un système pour télétransmettre les goûts et les odeurs. Le pilote d'un robot de téléprésence aurait ainsi pu participer à distance aux repas, qui sont des moments de convivialités importants. Or, au cours des prétests de ce système, les chercheurs se seraient rendu compte que la culture des sujets et leur posture corporelle au moment d'interagir entre eux avaient une influence sur leurs perceptions. Une étude était donc soit disant organisée pour élucider ce problème. Les sujets qui répondaient à l'avis de recherche étaient prévenus que l'expérience se déroulait en binôme, et qu'ils seraient testés soit sur le goût, soit sur l'odorat.

Figure 62 : Affiche utilisée pour recruter des sujets

4.1.2.2 Déroulement

À son arrivée, le sujet était accueilli par moi-même dans la chambre sourde, aménagée en salle d'attente. Il y rencontrait Véronique Aubergé, qui se faisait passer pour le second sujet de l'expérience. Pendant que je rappelais le contexte de l'expérience, elle expliquait qu'elle s'était portée volontaire car elle était confrontée au même problème dans son métier : elle prétendait être un nez professionnelle, travaillant à Uriage, célèbre entreprise de cosmétique[19]. Selon elle, Uriage tentait de s'implanter au Japon, et les goûts des consommateurs nippons différaient sensiblement des goûts des français. Après quelques minutes de discussion libre, les sujets remplissaient un formulaire de renseignements et un formulaire de consentement. Puis, nous

[19] Notons que cette manière d'introduire Aubergé comme une experte de l'odorat permettait de justifier le fait qu'au cours de l'expérience elle alterne entre des moments de grande confiance et de grand doute.

passions sur la plateforme expérimentale, ou j'expliquais plus en détail le protocole de l'expérience.

L'objectif était d'évaluer si les perceptions olfactives et gustatives pouvaient être influencées socialement. Les sujets étaient donc placés dans différentes situation d'interaction : soit face à face, dos à dos, ou côte à côte. Comme Aubergé était nez professionnel, elle était assignée au rôle de goûteuse, car ses performances olfactives auraient été bien supérieures à la moyenne. Elle devait se déplacer dans la pièce en suivant l'ordre noté sur une feuille de papier, et goûter aux pilules disposées dans différents gobelets en plastique. Après avoir goûté une pilule, elle devait annoncée le mot à voix haute, puis une boîte à odeur correspondant à la même essence serait donnée à l'autre sujet, qui devait annoncer à voix haute l'odeur qu'il avait reconnu. Les sujets avaient alors la possibilité de discuter de leur choix, et d'éventuellement le modifier. Ils étaient équipés de microphones, afin que l'on puisse conserver une trace de leurs discussions.

Cependant, je précisais que l'on ne voulait pas que les sujets puissent être influencés par les émotions ressenties par leurs binômes : en particulier, les expressions de plaisir ou de dégoût apparaissant sur leur visage auraient pu influencer leurs réponses. Il fallait donc qu'ils ne puissent pas voir le visage de l'autre. Ainsi, le sujet chargé de respirer les odeurs devait obligatoirement porter un masque qui l'aveuglait et lui couvrait entièrement le visage (cf. Figure 63).

Figure 63 : Photo illustrative du déroulement de l'expérience

Malheureusement, le sujet aveuglé risquait de perdre le contact avec sa binôme, et de se contenter de donner ses réponses dans le vague. Pour s'assurer qu'il reste bien concentré sur l'interaction, on lui demandait de la localiser dans l'espace. Avant de donner une réponse, il devait donc annoncer dans quelle direction elle se trouvait, à quelle distance et si elle était tournée vers lui ou vers le mur. Ils étaient prévenus que pendant ses déplacements, un extrait musical serait joué pour passer le temps et couvrir ses bruits de pas.

Le sujet était alors installé, et le test pouvait commencer. Il durait entre 40 min et 1 heure, selon le temps de réponse des sujets. Pour chaque passation, 64 mots-clés étaient prononcés, et 64 boîtes à odeurs données à sentir aux sujets.

Une fois le test terminé, le sujet pouvait ôter son masque. Je demandais ce qu'ils avaient pensé de l'expérience, puis j'annonçais que je ne leur avais pas tout dit, et leur demandais de deviner quel était le vrai but de l'expérience. Après que le sujet ait fait quelques propositions, Aubergé dévoilait son rôle. Nous expliquions alors les raisons de cette mise en scène, et ce que nous voulions tester. Une fois bien informé, le sujet avait la possibilité de refuser l'utilisation de ses données. S'il acceptait malgré tout qu'elles soient utilisées, il signait un second formulaire de consentement.

4.1.2.3 Dispositif

L'expérience se déroulait dans les anciens locaux de Domus, plus précisément sur le plateau expérimental. Il s'agissait d'une salle de 7.1 x 8.6 m, particulièrement réverbérante[20] du fait de son haut plafond technique, de son sol et de ses murs nus et de sa baie vitrée (cf. Figure 64). Pour cette expérience, la majorité des meubles était rassemblée dans un coin de la pièce, derrière des paravents, de manière à former un espace carré de 7.1 x 7.1 m, ce qui correspond à un carré de 10 m de diagonale.

Figure 64 : Photo du dispositif expérimental

La chaise du sujet était disposée au centre de la pièce, devant une table de hauteur variable, positionnée de manière à supporter deux haut-parleurs placés au niveau des oreilles. À la droite de cette chaise, une table basse portait les boîtes à odeurs utilisées dans l'expérience. Devant la table basse, une seconde chaise était disposée à mon intention.

Deux porte-documents avec les ordres de passage étaient disposés dans la pièce. Le premier contenait l'ordre de passage des boîtes à odeurs (cf. Figure 65). Il indiquait la couleur et le numéro des boîtes, ainsi que les déplacements prévus pour la locutrice, afin de me permettre de les vérifier au cours de l'expérience. Les cases grisées dans l'ordre de passage correspondent à l'orientation dos au sujet (l'orientation face au sujet étant l'orientation par défaut). Le second porte-document contenait l'ordre de passage d'Aubergé. Sur la première page n'apparaissait que les positions où elle devait se rendre, ainsi que les numéros des gobelets (cf. Figure 67). Derrière cette feuille était glissé le vrai ordre de passage, indiquant également le socio-affect, l'intensité et le mot à prononcer (cf. Figure 66).

Les voix des deux participants étaient enregistrées à l'aide de microphones-HF Sennheiser HSP4, et d'une carte son UR22mkII.

[20] Le temps de réverbération mesuré était de l'ordre de 0.8 s.

Rouge

1	derrière	milieu	12	
2	gauche	proche	7	
3	droite	proche	6	
4	avant	milieu	18	
5	derrière	milieu	9	

Figure 65 : Extrait de mon ordre de passage

	Direction	Distance	Gobelet	
1	derrière	milieu	G1	
2	gauche	proche	D1	
3	droite	proche	B1	
4	avant	milieu	I1	
5	derrière	milieu	G2	

Figure 67 : Extrait de l'ordre de passage montré au sujet

1	derrière	milieu	doute POL	faible	cassis	
2	gauche	proche	doute POL	fort	rose	
3	droite	proche	confiance AUT	fort	lavande	
4	avant	milieu	confiance AUT	fort	banane	
5	derrière	milieu	confiance AUT	faible	pomme	

Figure 66 : Extrait de l'ordre de passage donné à Aubergé

4.1.3 Analyse des enregistrements

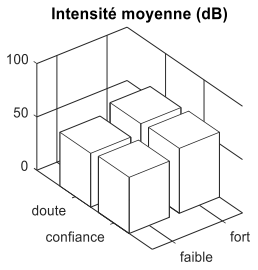
Cette partie concerne l'analyse a posteriori des mots-clés prononcés pendant l'expérience. Il s'agit de vérifier que les consignes données à V. Aubergé ont bien été respectées. Après un premier découpage grossier des enregistrements sur Audacity, chaque mot-clé a été redécoupé à la main en utilisant le logiciel Praat.

4.1.3.1 Intensité moyenne

Les mesures d'intensité moyenne sont rapportées dans le Tableau 5. On constate que les mots-clés de faible intensité diffèrent significativement de ceux d'intensité forte (- 8,4 dB en moyenne). Cela signifie que la consigne en intensité a été bien respectée. L'écart-type est relativement élevé pour chaque classe, mais cela s'explique en partie par le fait que l'intensité varie d'un mot-clé à l'autre. L'intensité moyenne varie également en fonction du socio-affect : le doute faible est 2,1 dB plus faible que la confiance faible, et la confiance forte 1,7 dB plus forte que le doute fort.

Tableau 5 : Intensité des mots-clés pour chaque classe

Classe	Moyenne (dB)	Écart-type (dB)	Nombre de stimuli
doute faible	50,5	4,7	160
doute fort	59,2	4,8	159
confiance faible	52,6	4,6	160
confiance fort	60,8	4,1	161

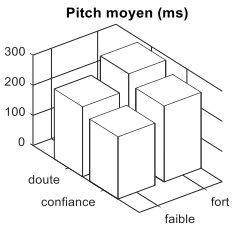


4.1.3.2 Fréquence fondamentale

La fréquence fondamentale varie également en fonction de l'intensité et du socio-affect (cf. Tableau 6) : elle est plus élevée lorsque l'intensité est forte, et plus faible pour la confiance que pour le doute.

Tableau 6 : Fréquence fondamentale des mots-clés pour chaque classe

Classe	Moyenne (Hz)	Écart-type (Hz)	Nombre de stimuli
doute faible	232	24	160
doute fort	285	32	159
confiance faible	207	19	160
confiance fort	252	25	161



4.1.3.3 Évolution temporelle de l'intensité et du pitch

La Figure 68 représente le tracé dynamique de l'intensité et du pitch. Les courbes obtenues sont cohérentes avec les mesures moyennes précédentes : en particulier la courbe d'intensité faible est plus basse sur l'échelle des ordonnées que celle d'intensité forte. L'allure des courbes est également très caractéristique : on observe un pic d'intensité pour la confiance, beaucoup plus étalé dans le cas du doute ; de plus, le pitch est ascendant dans le cas du doute, et descendant dans le cas de la confiance.

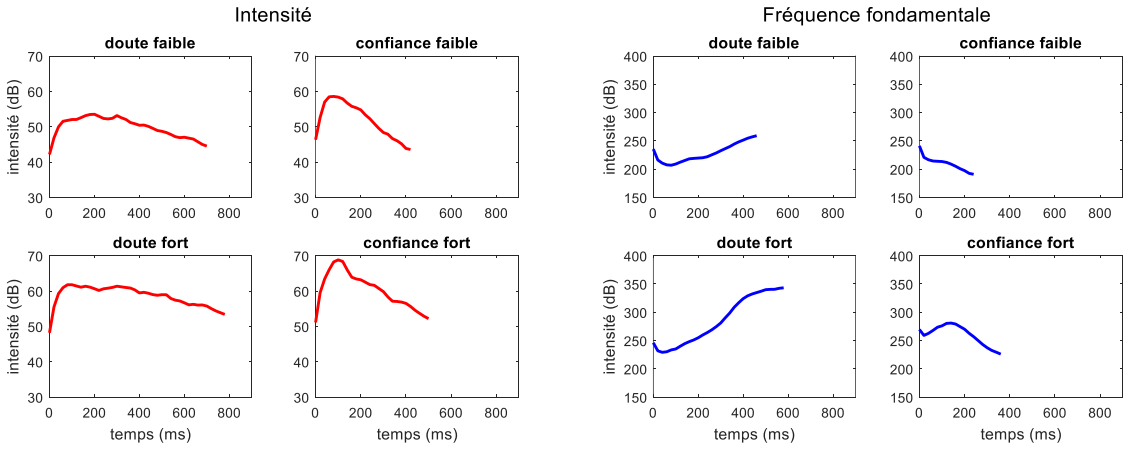


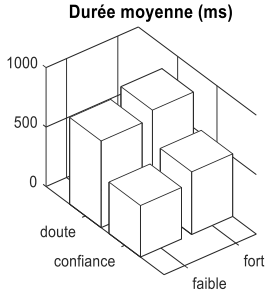
Figure 68 : Courbes prosodiques moyennes

4.1.3.4 Durée

La durée des mots-clés varie significativement d'une classe à l'autre, comme le rapport le Tableau 7. En moyenne, le doute est plus long que la confiance d'environ 280 ms. L'écart type est encore une fois élevé, car le nombre de syllabes varie d'un mot à l'autre. Elle est également influencée par la consigne d'intensité : ainsi, le doute fort est plus long que le doute faible d'environ 70 ms, et la confiance faible est plus courte que la confiance forte d'environ 15 ms. Ainsi, pour conserver la douceur du doute, tout en augmentant son intensité, Aubergé a eu

Tableau 7 : Durée des mots-clés pour chaque classe

Classe	Moyenne (ms)	Écart-type (ms)	Nombre de stimuli
doute faible	729	176	160
doute fort	798	228	159
confiance faible	433	126	160
confiance fort	526	139	161



besoin d'allonger son geste. À l'inverse, pour produire un signal agressif, mais de faible intensité, elle a raccourci son geste.

4.1.3.5 Qualité de voix

Par ailleurs, Véronique Aubergé a réalisé un étiquetage des mots-clés en termes de qualité de voix, à l'aide d'une interface web lui permettant d'écouter plusieurs fois chaque enregistrement, et de sélectionner la ou les qualités de voix associées. Pour réaliser cette interface, ainsi que nos autres tests en ligne, nous nous sommes inspirées du travail de Clarisse Bayol, réalisé pendant son stage de Master 2 (2016), et utilisé pour l'étude de (Magnani et al. 2017).

On constate que chaque socio-affect est associé à une qualité de voix : *lax* / *breathy* pour le doute, et *tense* pour la confiance. Lorsque l'intensité est cohérente avec le socio-affect (doute faible et confiance fort), le pourcentage de mots-clés réalisés avec la bonne qualité de voix s'élève à près de 95%, tandis que le pourcentage de mots-clés réalisés avec la qualité de voix opposée est anecdotique, de l'ordre de 1%. Pour les classes mixtes (doute fort et confiance faible), la qualité de voix est moins régulière : environ la moitié des mots-clés ont une qualité de voix modale, et 8 à 30 % des mots-clés ont une qualité de voix opposée à celle recherchée.

Tableau 8 : Qualité de voix des mots-clés pour chaque classe (%)

Classe	Lax	Breathy	Modal	Tense	Creaky	Harsh
doute faible	**93,7**	**95,0**	7,5	1,3	20,7	25,1
doute fort	**71,5**	**82,3**	**59,5**	27,8	10,1	13,3
confiance faible	8,7	12,5	**38,1**	**87,5**	7,5	25,6
confiance fort	0,6	0,6	15,1	**97,5**	8,8	24,5

Ces analyses confirment le ressenti de la locutrice pendant les expériences : les mots des catégories « doute faible » et « confiance fort » sont plus faciles à produire que les autres, car les deux consignes sont cohérentes. Pour produire les deux autres catégories, elle a dû exagérer certains indices prosodiques : durée et fréquence fondamentale. De plus, la production d'une intensité incohérente se fait au détriment de la qualité de voix.

4.1.3.6 Écoute des enregistrements

À la suite de la première expérience, nous avons mis en place un test en ligne, afin de vérifier si les socio-affects des 640 mots-clés enregistrés étaient bien reconnaissables. Le site web était hébergé sur les serveurs de l'Imag, l'institut d'Informatique et de Mathématiques Appliquées de Grenoble. Pour concevoir l'interface de test, nous nous sommes inspirés du travail effectué par Clarisse Bayol pendant son stage de Master 2 (2016).

Au cours du test, les participants devaient écouter une liste d'enregistrements, et indiquer pour chacun d'entre eux le type de voix reconnu (petit doute poli / gros doute poli / petite confiance autoritaire / grosse confiance autoritaire) et leur niveau de certitude (pas sûr / sûr). Notons qu'ils étaient interrogés non pas sur l'intensité acoustique, mais sur l'intensité du socio-affect produit.

Le nombre de participants s'élève à 32 : il s'agit de 18 femmes et 14 hommes dont la moyenne d'âge est de 34 ± 14 ans. Dans un premier temps, chaque sujet écoutait dans un ordre aléatoire 1/3 des 640 enregistrements, soit 213 ou 214 mots-clés. Nous avons rapidement réduit ce nombre à 160, car certains sujets mettaient beaucoup plus de temps à répondre que les 20 min initialement estimées. Chaque enregistrement a été réécouté par 5 à 11 auditeurs différents.

Le Tableau 9 représente la matrice de confusion entre la consigne donnée à la locutrice, et l'étiquette majoritaire obtenu par chaque enregistrement. On constate que pour 90 % des mots-clés, l'étiquette majoritaire correspond bien à la consigne reçue par la locutrice[21].

Tableau 9 : Matrice de confusion pour la classification des 640 mots-clés de l'expérience 1

		Consigne de production	
		doute	confiance
Etiquette majoritaire	doute	274 **86 %**	18 6 %
	confiance	45 14 %	303 **94 %**

Lorsqu'on s'intéresse au détail des résultats, on constate que le doute était plus difficilement reconnu que la confiance. Ainsi, la Figure 69 représente deux matrices contenant le décompte des réponses des sujets en fonction de l'étiquette de production. La matrice verte en haut de la figure correspond aux cas où les sujets ont indiqués qu'ils étaient sûrs de leur réponse ; tandis que la matrice rouge en bas de la figure correspond aux cas où les sujets ont indiqués qu'ils n'étaient pas sûrs de leur réponse. On constate que les sujets étaient particulièrement certains de reconnaître une « grosse confiance » lorsque le type de voix utilisé était la « confiance forte » ; au contraire, ils étaient bien moins sûrs d'eux au moment d'identifier le doute ou la

[21] Le doute étant moins bien reconnu en moyenne que la confiance, c'est l'étiquette « doute » qui a été retenue en cas d'égalité.

« petite confiance ». Certains sujets ont d'ailleurs précisé en commentaire à la fin de l'expérience qu'ils n'avaient pas bien compris ce à quoi correspondait un « doute poli ».

Par ailleurs, on constate que le classement en terme d'intensité n'est pas très clair (cf. Tableau 10). A priori, on s'attendait à ce que les mots-clés dont l'intensité est cohérente avec le socio-affect exprimé soient perçus comme des « gros doute » ou des « grosse confiance ». En pratique ce n'est pas le cas : dans le cas de la « confiance faible » et du « doute fort », les sujets ont privilégié respectivement les étiquettes « grosse confiance » et « gros doute ». Il ne semble pas que cela soit dû à une mauvaise interprétation des consignes : en effet, les sujets semblent avoir bien compris que les catégories extrêmes étaient le « gros doute » et la « grosse confiance », puisque ce sont les catégories les moins choisies lorsque le socio-affect est respectivement la « confiance » et le « doute ».

Tableau 10 : Pourcentage de réponses dans chaque catégorie

		Consigne de production			
		doute faible	doute fort	confiance faible	confiance forte
Etiquette de perception	grosse confiance	29 2 %	137 10 %	606 **45 %**	1097 **81 %**
	petite confiance	357 27 %	267 20 %	468 34 %	139 10 %
	petit doute	459 34 %	421 31 %	233 17 %	86 6 %
	gros doute	493 **37 %**	517 **39 %**	50 4 %	29 2 %

Les socio-affects exprimés au cours de l'expérience sont donc bien reconnaissables ; du moins à condition que les sujets soient informés de leur existence.

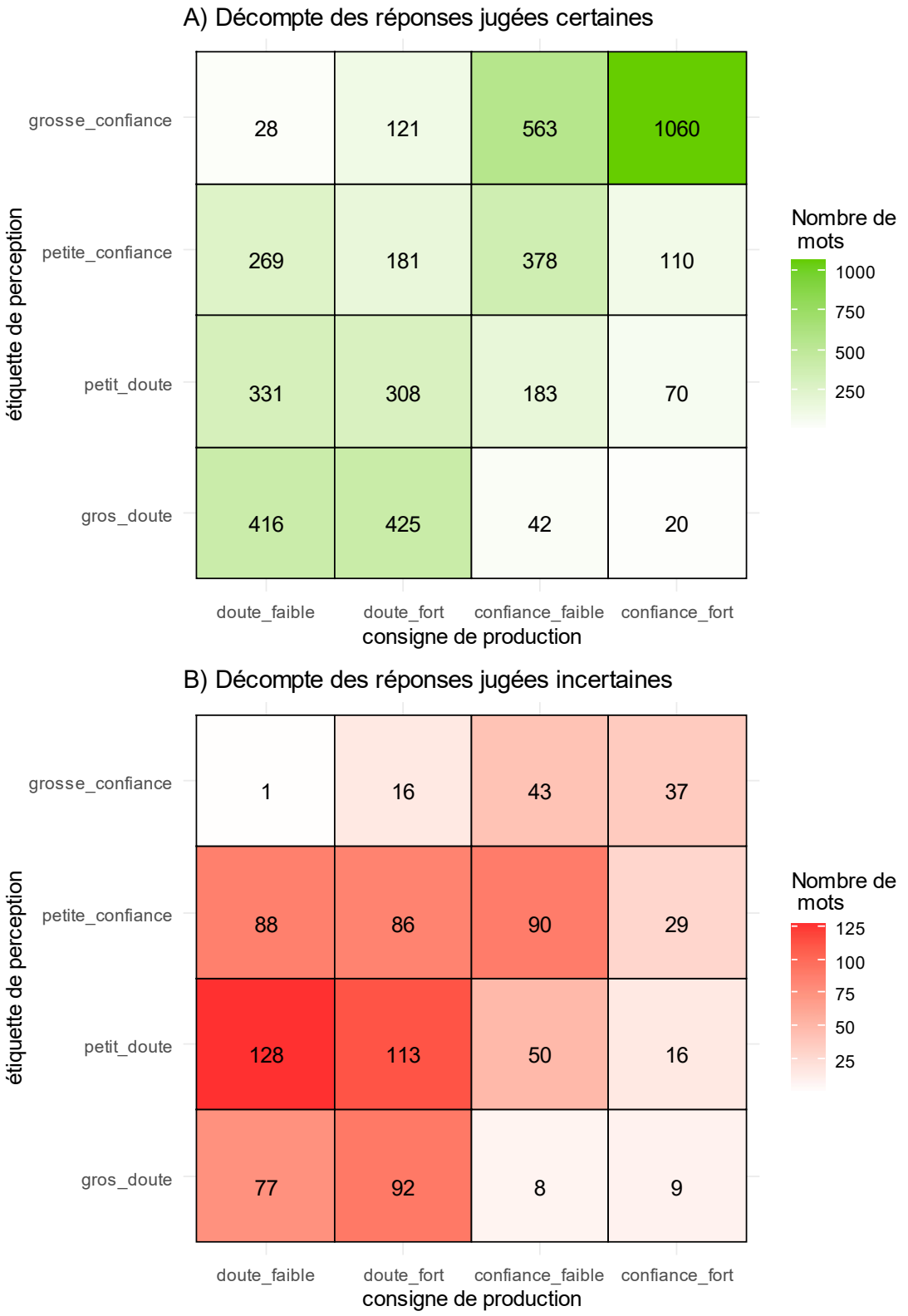


Figure 69 : Perception du socio-affect en fonction de la consigne de production

4.1.4 Analyse statistique des résultats

Nous allons à présent passer à l'analyse des résultats. L'expérience durait environ 1h30 pour 64 mots-clés, dont chacun correspondait à une combinaison possible des deux facteurs décrivant le type de voix utilisé (socio-affect et intensité) et des trois variables spatiales (direction, distance et orientation). Elle a été passée par 10 sujets (3 femmes et 7 hommes), la plupart étudiants à l'université Grenoble Alpes.

4.1.4.1 Premières observations

Commençons par comparer les réponses des sujets aux positions spatiales de la locutrice, sans tenir compte des facteurs sous-jacents. Ces premiers résultats sont rapportés dans les matrices de confusion ci-dessous, sous la forme de pourcentage. En haut à gauche de chaque case apparaît le nombre exact de données recueillies dans chaque catégorie.

La variable la plus simple à percevoir était la **direction** (cf. Tableau 12). Ainsi, 91% des réponses des sujets sont justes. Parmi les erreurs, on compte principalement des confusions sémantiques gauche – droite, et des confusions acoustiques avant – derrière. Notons que 90 % des confusions acoustiques se font dans le sens derrière perçu comme avant. Les sujets auraient donc tendance à percevoir leur interlocuteur devant eux.

La **distance** était la variable la plus difficile à percevoir (cf. Tableau 13). Ainsi, seuls 59 % des réponses des sujets sont justes. La distance « proche » est la mieux reconnue, avec un taux de reconnaissance de 83 %. En revanche, les sujets ont beaucoup hésité entre les distances « milieu » et « loin », pour lesquelles le taux de reconnaissance tombe à 50 %.

En ce qui concerne, la perception de **l'orientation**, les sujets ont donné la bonne réponse dans 79 % des cas (cf. Tableau 14). En moyenne, l'orientation de dos était mieux perçue que l'orientation de face.

Dans la suite, on s'intéressera uniquement au taux de reconnaissance, c'est-à-dire au pourcentage de réponses justes pour une variable donnée. Notons que le taux de reconnaissance moyen varie d'un sujet à l'autre (cf. Tableau 11).

Tableau 11 : Taux de reconnaissance sujet par sujet (%), accompagnés de la moyenne μ et de l'écart-type σ

Sujet	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	***μ***	***σ***
direction	78,1	87,5	100	90,6	81,2	95,3	87,5	100	95,3	98,4	**91,4**	**7,7**
distance	67,2	67,9	54,7	53,1	43,7	67,2	NA	54,0	57,8	62,5	**58,7**	**8,2**
orientation	83,8	71,9	70,3	76,6	68,8	87,5	83,9	92,2	68,7	92,2	**79,5**	**9,4**

Tableau 12 : Matrice de confusion pour la perception de la direction

direction		**Production**			
		avant	gauche	droite	derrière
Perception	avant	156 **97 %**	0 0 %	1 1 %	26 16 %
	gauche	0 0 %	152 **94 %**	6 4 %	0 0 %
	droite	1 1 %	6 4 %	146 **92 %**	3 2 %
	derrière	3 2 %	3 2 %	6 4 %	131 **82 %**

Tableau 13 : Matrice de confusion pour la perception de la distance

distance		**Production**		
		proche	milieu	loin
Perception	proche	116 **83 %**	35 12 %	13 9 %
	milieu	20 14 %	136 **48 %**	51 36 %
	loin	4 3 %	111 39 %	78 **55 %**

Tableau 14 : Matrice de confusion pour la perception de l'orientation

orientation		**Production**	
		face	dos
Perception	face	371 **78 %**	24 15 %
	dos	107 22 %	136 **85 %**

4.1.4.2 Influence du socio-affect et de l'intensité

On s'intéresse à présent à la variation du taux de reconnaissance, en fonction du socio-affect et de l'intensité. C'est là l'objectif principal de l'étude.

Chaque mot-clé prononcé par la locutrice appartenait à une des quatre catégories suivantes : « doute faible », « doute fort », « confiance faible » ou « confiance forte ». Il est donc possible de calculer un taux de reconnaissance pour chaque sujet pour chaque catégorie, et de voir si le taux varie significativement d'une catégorie à l'autre. Les graphiques correspondant apparaissent en Figure 70. Ils sont accompagnés d'une projection, qui isole les taux de reconnaissance en fonction du socio-affect d'un côté, et en fonction de l'intensité de l'autre. Sous chaque projection apparaît une mesure de la valeur-p, obtenue à l'aide d'une Anova. La valeur-p représente la probabilité d'obtenir nos résultats, en supposant que les variables étudiées n'ont aucun effet sur les données (hypothèse nulle). Plus la valeur-p est faible, plus cette hypothèse nulle est improbable : en dessous d'un seuil fixé arbitrairement, le résultat est qualifié de significatif, ce qui signifie que la variable étudiée a très probablement un effet sur les données. Chaque valeur-p est accompagnée d'un symbole (« . » , « * », « ** » ou « *** »), en fonction du seuil de significativité passé (0,1 ; 0,05 ; 0,01 ; 0,001).

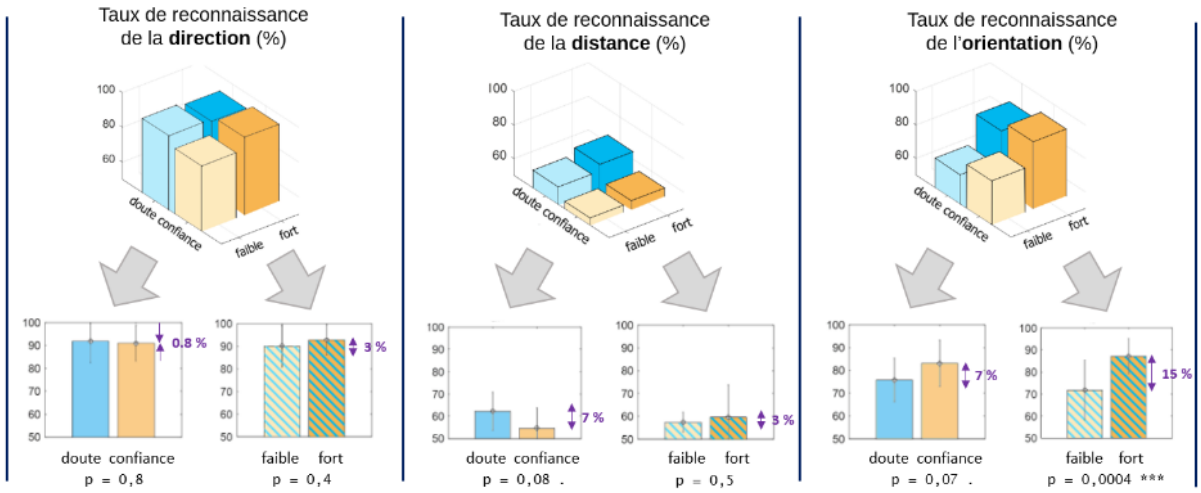


Figure 70 : Taux de reconnaissance en fonction des quatre catégories (socio-affect × intensité)[22]

En moyenne, le taux de reconnaissance de la **direction** ne varie pas significativement en fonction du socio-affect et de l'intensité. Deux interprétations sont possibles. Soit le socio-affect et l'intensité n'ont pas d'impact sur la perception de la direction, soit cet effet est trop faible pour pouvoir être observé compte tenu du nombre de données recueilli. En effet, on compte seulement 55 erreurs sur la perception de la direction pour les 10 sujets, ce qui signifie qu'en moyenne chaque sujet s'est trompé entre 5 et 6 fois dans ses réponses. Calculer les taux de

[22] Les valeurs-p sont légèrement différentes de celles présentées dans l'article (Davat et al. 2020), car ici, le modèle étudié est directement « taux ~ socio_affect*intensite », et non « taux ~ socio_affect » d'un côté et « taux ~ intensite » de l'autre.

reconnaissance pour chaque catégorie revient à classer ces 5 - 6 erreurs en 4 catégories. Les tendances observées d'un sujet à l'autre sont donc contradictoires : par exemple, environ la moitié des sujets obtient de meilleurs taux de reconnaissance lorsque l'intensité est forte, et l'autre moitié lorsque l'intensité est faible. Cependant, nous optons plutôt pour la première interprétation, car nous n'avons rien trouvé dans la littérature qui permette de penser que la perception de la direction pourrait être influencée par le type de voix utilisé.

Pour la **distance**, le nombre de données disponibles est plus important (234 erreurs pour 9 sujets[23]). L'intensité n'a aucun effet significatif sur le taux de reconnaissance, ni sur la répartition des réponses des sujets (cf. Tableau 15). Ce résultat est assez étonnant, car l'intensité est un indice important pour la perception de la distance : on pourrait s'attendre à ce que les mots d'intensité faible soient perçus plus loin que ceux d'intensité fort, puisque l'intensité diminue avec la distance. Au contraire, les sujets pourraient également interpréter l'augmentation d'intensité comme liée à un éloignement : la locutrice parlant plus fort pour compenser la distance qui la sépare de son auditeur. Ici, on constate simplement une augmentation faible et non significative du nombre de mots-clés perçus à distance proche lorsque l'intensité est forte.

Tableau 15 : Décompte des distances perçues par les sujets en fonction de l'intensité

intensité	proche	milieu	loin
faible	77	107	100
fort	87	100	93

En moyenne, le taux de reconnaissance était légèrement plus faible pour la « confiance » que pour le « doute » (- 7%), mais cette variation est peu significative (valeur-p : 0,08). Lorsqu'on regarde la répartition des réponses, on constate une augmentation faible et non significative du nombre de mots-clés perçus à distance proche lorsque le socio-affect est la « confiance » (cf. Tableau 16).

Tableau 16 : Décompte des distances perçues par les sujets en fonction du socio-affect

socio-affect	proche	milieu	loin
doute	79	108	97
confiance	85	99	96

Il n'y a pas non plus d'effet croisé important entre les facteurs intensité et socio-affect : ainsi, on trouve à peu près autant de réponses dans chaque catégorie (cf. Tableau 17).

Tableau 17 : Décompte des distances perçues par les sujets en fonction du socio-affect et de l'intensité

socio-affect + intensité	proche	milieu	loin
doute faible	36	56	52
doute fort	43	52	45
confiance faible	41	51	48
confiance forte	44	48	48

[23] Les données sur la distance d'un des sujets n'ont pas été prises en compte, car il avait compris qu'il n'y avait que deux distances : proche et loin.

Les résultats concernant l'**orientation** sont plus marqués, bien que le nombre d'erreurs soit plus faible que pour la distance (seulement 131 pour les 10 sujets). On observe ainsi un effet significatif de l'intensité : le taux de reconnaissance est près de 15% plus élevé lorsque l'intensité est forte (valeur-p : 0,0004). En calculant le risque relatif à partir des données du Tableau 18, on constate que les sujets ont 1,8 fois plus de chances de percevoir la locutrice de dos lorsque son intensité est faible (IC95%[24] : [1,5 – 2,2]). Ce résultat est cohérent avec l'expérience acoustique des sujets : lorsque quelqu'un nous tourne le dos, l'intensité que nous percevons diminue.

Tableau 18 : Décompte des orientations perçues par les sujets en fonction de l'intensité

intensité	face	dos
faible	163	156
forte	232	87

De la même manière, le taux de reconnaissance est un peu plus élevé pour la « confiance » que pour le « doute » (+ 7%). En calculant le risque relatif à partir des données du Tableau 19, on constate que les sujets ont 1,2 fois plus de chances de percevoir la locutrice de dos lorsqu'elle exprime du « doute » (IC95% : [1,0 – 1,5]). Ce résultat est cohérent avec l'intensité intrinsèque des deux socio-affects.

Tableau 19 : Décompte des orientations perçues par les sujets en fonction du socio-affect

socio-affect	face	dos
doute	184	134
confiance	211	109

Enfin, en considérant uniquement les deux catégories extrêmes à partir des données du Tableau 20, on constate que les sujets ont 2,2 fois plus de chances de percevoir la locutrice de dos lorsqu'elle exprime un « doute faible » plutôt que de la « confiance forte » (IC95% : [1,6 – 3,0]).

Tableau 20 : Décompte des orientations perçues par les sujets en fonction du socio-affect et de l'intensité

socio-affect + intensité	face	dos
doute faible	74	86
doute fort	110	48
confiance faible	89	70
confiance forte	122	39

[24] Intervalle de confiance à 95% : le risque relatif « réel » a 95% de chance de se trouver dans cet intervalle. Comme il ne contient pas la valeur 1, on conclut à nouveau que l'effet est significatif.

En résumé, il n'y a donc pas d'effet marqué du type de voix utilisé sur la perception de la direction. Pour la perception de la distance, on note un léger effet du socio-affect : le taux de reconnaissance étant légèrement meilleur pour le doute que pour la confiance. Le résultat le plus significatif concerne la perception de l'orientation en fonction de l'intensité : les sujets avaient tendance à percevoir la locutrice de dos lorsque son intensité était faible, ou que le socio-affect utilisé était le doute. Cela est peut-être dû au fait que les sujets ont eu l'occasion de discuter dans la salle avant le début de l'expérience : ils ont donc pu constater que le fait de se déplacer dans la pièce n'engendrait pas d'importantes variations d'intensité, et les ont donc attribué directement à la locutrice. Comme ils ignoraient que la locutrice contrôlait sciemment son intensité de voix, ils ont pu conclure qu'elle leur tournait le dos lorsque son intensité était plus faible à cause du type de voix utilisé.

4.1.4.3 Influence des variables spatiales

De façon secondaire, on peut s'intéresser à la manière dont les réponses des sujets se répartissent en fonction des variables spatiales ; sachant que pour chaque mot-clé, ils avaient le choix entre 4 (directions) × 3 (distances) × 2 (orientations) = 24 combinaisons possibles (cf. Figure 71).

La diagonale correspond aux réponses justes. Quatre blocs ressortent principalement, correspondant aux quatre directions étudiées. À l'intérieur de ces quatre blocs, la répartition des réponses est similaire ; il semble donc que la direction n'ait pas eu d'effet sur la perception des sujets. De façon prévisible compte tenu des matrices de confusion précédentes, ce sont les mots-clés produits à la distance proche qui sont les mieux reconnus. On remarque également que si les mots-clés produits de dos sont majoritairement perçus de dos, leur distance est souvent surestimée (cf. Tableau 21). Ainsi, bien que 100 % des mots-clés « de dos » soient produits à la distance « milieu », 53 % d'entre eux sont perçus « loin ». Au contraire, la répartition des mots-clés « de face » est proche de la répartition initiale : environ 1/3 pour chaque distance.

Tableau 21 : Perception de la distance en fonction de l'orientation produite ou perçue

Distance perçue	**proche**	**milieu**	**loin**
Sons perçus de face	36 %	33 %	30 %
Sons produits de face	37 %	35 %	28 %
Sons perçus de dos	18 %	42 %	40 %
Sons produits de dos	4 %	43 %	53 %

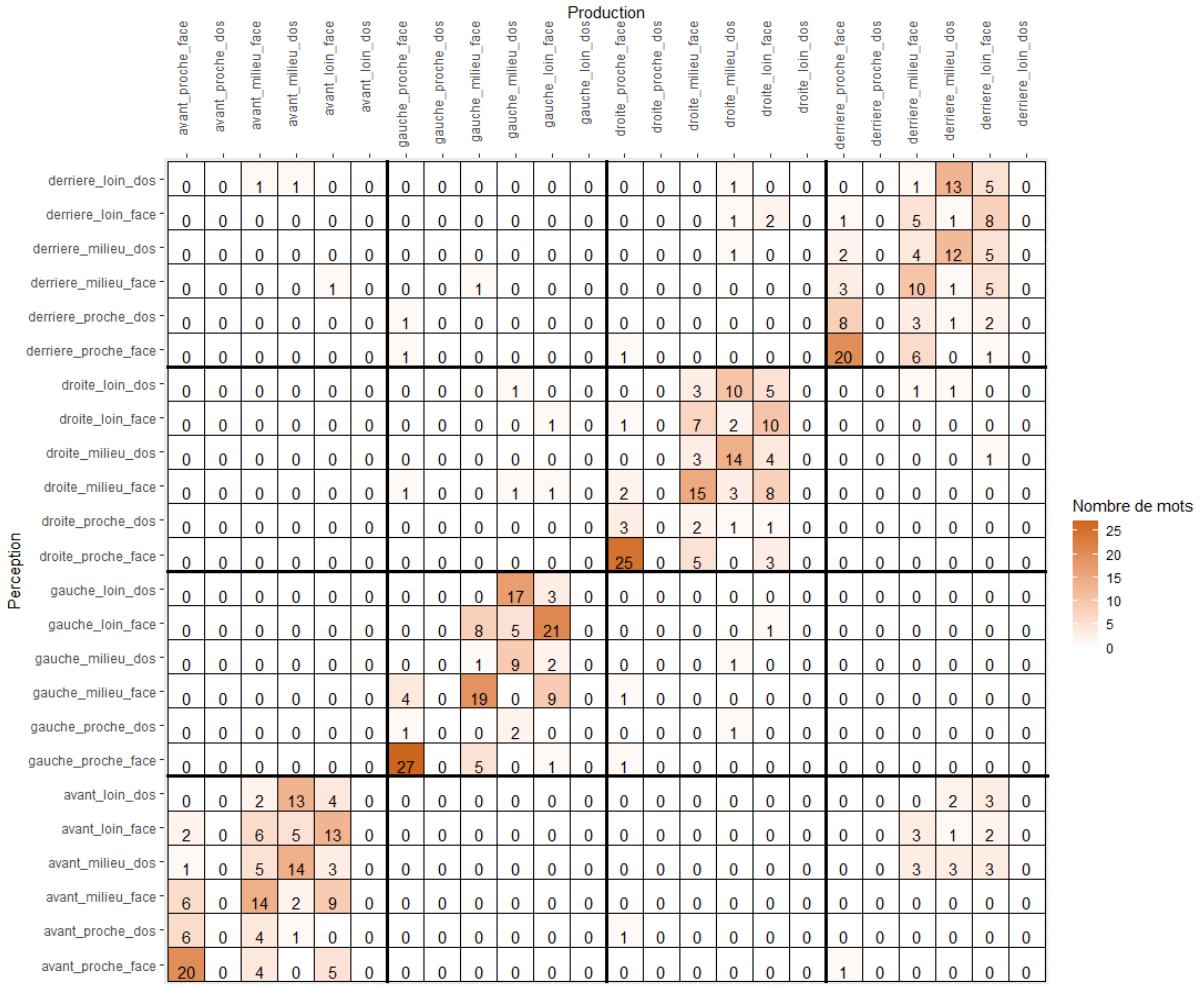


Figure 71 : Répartition des réponses des sujets en fonction des variables spatiales

4.1.4.4 Résumé

Cette première expérience nous a permis d'étudier la manière dont le type de voix pouvait influencer la perception de la position spatiale d'une locutrice. Ainsi, il s'agissait principalement d'étudier l'effet de deux facteurs (socio-affect et intensité) sur la perception de trois variables (direction, distance et orientation). Nos résultats permettent de nous faire une idée du sens dans lequel le type de voix peut influencer la perception ; ainsi que de l'importance de ses effets.

Concernant la perception de la direction, on retrouve un résultat connu de la psychoacoustique : les sujets identifient très bien la direction générale de leur interlocutrice, bien qu'il existe des confusions avant/arrière. Comme prévu, on n'observe pas d'effet du type de voix sur la perception de la direction ; ni d'effet de la direction sur la perception des autres variables. En revanche, l'intensité a bien un effet, en particulier sur la perception de l'orientation. Ainsi, lorsque l'intensité était faible, les sujets ont eu tendance à croire que la locutrice leur tournait le dos. De plus, les sons perçus de dos étaient souvent perçus plus loin que leur véritable distance. L'effet du socio-affect est plus discret, mais cohérent avec les variations d'intensité :

le doute étant intrinsèquement plus faible que la confiance, comme l'ont montré les analyses des enregistrements.

Du fait du prétexte expérimental, cette expérience était longue et éprouvante, tant pour les sujets, qui devaient respirer plusieurs dizaines de boîtes à odeurs, que pour les expérimentateurs, qui devaient jouer la comédie. Pour des raisons d'emploi du temps, seuls 10 sujets ont finalement passé l'expérience. De plus, le nombre de données recueillies pour chaque sujet est faible compte tenu du nombre de facteurs et de variables étudiées. Malgré tout, des résultats significatifs ont pu être mis en évidence. Par la suite, nous avons cherché à les approfondir à travers une seconde expérience.

4.2 Reproduction de l'expérience sans prétexte

Pour compléter nos interprétations précédentes, une version simplifiée de l'expérience a été réalisée. Cette fois, pas de mise en scène : les sujets n'avaient pas à respirer de boîtes à odeurs, et savaient que l'expérience portait uniquement sur leur perception spatiale. De plus, ils étaient mis au courant des variations en intensité et en socio-affect, et devaient essayer de les reconnaître. L'objectif était ainsi de comparer les résultats obtenus dans deux conditions : avec ou sans prétexte. On s'attend à ce que lorsque les sujets prêtent attention au type de voix utilisé, celui-ci n'ait plus d'effet sur leur perception.

Cette expérience était beaucoup plus courte, et ne durait qu'entre 20 et 30 minutes. Elle a été passée par 8 sujets (4 femmes et 54 hommes). Dans l'idéal, on aurait souhaité faire revenir les sujets de la première expérience, malheureusement ils n'étaient pas disponibles au moment des passations ou n'ont pas répondu à nos sollicitations.

4.2.1 Résultats

Les résultats de cette seconde expérience sont présentés en Figure 72, en vis-à-vis des résultats de la première expérience. Cette fois, les taux de reconnaissance obtenus ne varient pas en fonction de la classe étudiée. Ils sont également plus élevés en moyenne et moins variables d'un sujet à l'autre (cf. Tableau 22).

Tableau 22 : Taux de reconnaissance pour les deux versions de l'expérience

Variable étudiée	**Taux de reconnaissances (moyenne ± écart-type)**	
	Avec prétexte	**Sans prétexte**
direction	91 ± 9 %	93 ± 8 %
orientation	79 ± 15 %	85 ± 10 %
distance	52 ± 21 %	58 ± 8 %

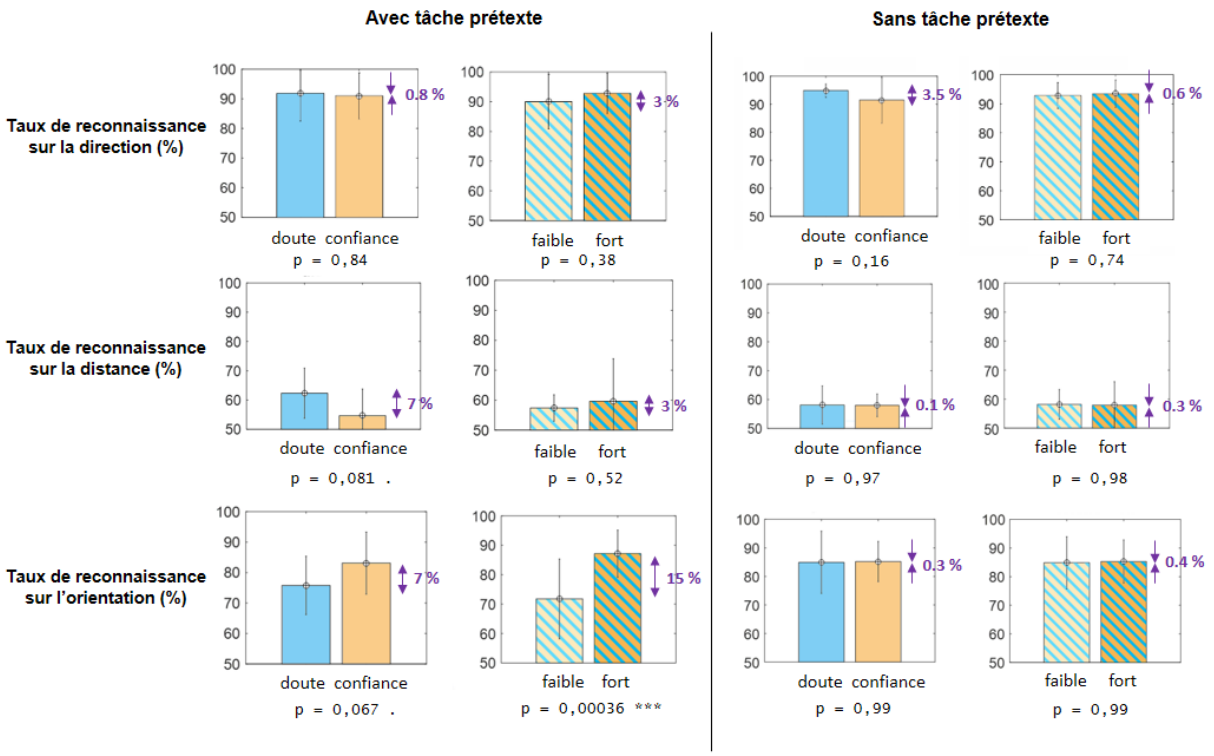


Figure 72 : Comparaison des résultats au test de localisation pour les deux versions de l'expérience

L'écart le plus frappant entre les deux expériences concerne la perception de l'orientation en fonction de l'intensité produite (cf. Tableau 23). Dans la première expérience, près de la moitié des mots-clés prononcés avec une intensité faible ont été perçus de dos. Cette tendance s'efface pour l'expérience sans prétexte : ainsi, on constate un report de 10% de la classe « faible perçu de dos », vers la classe « faible perçue de face ».

Tableau 23 : Pourcentage de réponses concernant l'orientation en fonction de l'intensité produire

Classe du mot-clé	Avec prétexte	Sans prétexte
faible perçu de face	25,5 %	35,3 %
fort perçu de face	36,4 %	36,3 %
faible perçu de dos	24,4 %	14,1 %
fort perçu de dos	13,6 %	14,3 %

Par ailleurs, le fait de demander aux sujets de reconnaître la classe des mots-clés permet d'en faire une validation perceptive (cf. **Erreur ! Source du renvoi introuvable.**). Les classes les mieux reconnues sont celles pour lesquelles le socio-affect est cohérent avec l'intensité : « doute faible » et « confiance forte ». En particulier, le « doute faible » n'est jamais perçu comme de la « confiance forte », et la « confiance forte » n'est jamais perçue comme du « doute ». Les classes ambiguës sont plus difficiles à reconnaître : les sujets se trompent dans la moitié des cas. En moyenne, le socio-affect (reconnu à 89%) était plus facile à reconnaître que l'intensité (reconnue à 76 %).

Tableau 24 : Matrice de confusion pour la perception de la classe du mot-clé dans l'expérience sans prétexte

Socio-affect + Intensité		**Production**			
		doute faible	doute fort	confiance faible	confiance forte
Perception	doute faible	53 **67 %**	15 **19 %**	8 **11 %**	0 **0 %**
	doute fort	18 **23 %**	37 **48 %**	1 **1 %**	0 **0 %**
	confiance faible	8 **10 %**	8 **10 %**	37 **50 %**	6 **7 %**
	confiance forte	0 **0 %**	17 **22 %**	28 **38 %**	81 **93 %**

Il semblerait donc que, dans cette deuxième expérience, la perception des variables spatiales n'ait pas été influencée par le type de voix utilisé. C'est exactement ce qu'on voulait vérifier. De plus, les sujets ont réussi à reconnaître le socio-affect et l'intensité utilisée dans la majorité des cas. On peut donc en déduire qu'en l'absence de prétexte, les sujets arrivent à se servir de cette information pour l'intégrer à leur perception spatiale.

4.2.2 Détails individuels

Pour compléter les interprétations précédentes, étudions les taux de reconnaissance individuels. A priori, une absence de variations peut s'observer dans deux cas :

- cas 1 : Effectivement les variables d'intérêt n'influencent pas les résultats : ainsi, la différence entre le taux de reconnaissance obtenu en condition A et celui obtenu en condition B est donc proche de 0 pour chaque sujet.
- cas 2 : Au contraire, les variables d'intérêt influencent les résultats, mais de façon différente d'un sujet à l'autre. Ce n'est qu'en moyennant les résultats qu'on n'observe une variation nulle.

Ces deux cas sont représentés en Figure 73 : les résultats des sujets sont toujours centrés autour de 0, mais leur dispersion varie. Seul le cas 1 permettrait de conclure définitivement que les variables d'intérêt n'influencent pas les résultats.

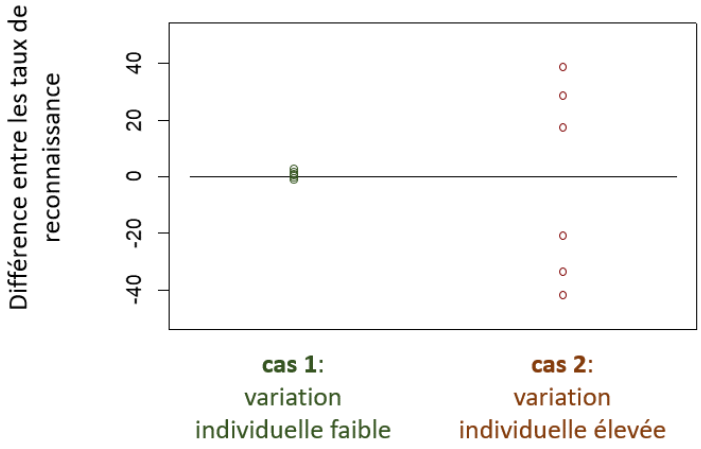


Figure 73 : Illustration de deux cas dans lesquels on n'observe pas de variation de taux de reconnaissance entre deux conditions

En pratique, les résultats obtenus se rapprochent plutôt du cas 2 : ils sont tout aussi dispersés pour l'expérience 1 que pour l'expérience 2 (cf. Figure 74). Or, on s'attendait plutôt à observer une diminution de la dispersion entre les deux expériences, étant donné la diminution de l'écart type observé au Tableau 22. En réalité, les sujets se trompent tout autant dans les deux expériences, mais pas de la même manière. Dans l'expérience 1, il arrive qu'ils se trompent tous dans la même direction : en particulier pour l'orientation leurs résultats sont systématiquement au-dessus de l'axe horizontal, car tous se sont moins souvent trompés pour la confiance que pour le doute. Dans l'expérience 2, les sujets sont également influencés à titre individuel par le socio-affect et l'intensité, puisque leurs résultats s'écartent de l'axe horizontal ; mais il n'y a pas de tendance globale : la moitié des sujets se trompe plutôt pour la confiance, et l'autre plutôt pour le doute.

On choisit de modéliser ces résultats par une distribution normale, car a priori, on s'attend à ce que les données se répartissent de façon symétrique autour d'une valeur moyenne. Selon le Tableau 25, les écart-types obtenus dans les deux expériences sont du même ordre de grandeur ; de plus, chaque point correspond à un seul sujet. Les hypothèses de normalité des distributions, d'homogénéité des variances, et d'indépendance des échantillons sont donc vérifiées, et on peut s'autoriser à utiliser une Anova. Les valeur-p obtenues correspondent à la probabilité que les deux jeux de données puissent être expliqués par une seule distribution normale. Elles sont rapportées dans le Tableau 26.

On constate donc que les conditions dans lesquelles les résultats des deux expériences sont significativement différents sont identiques aux conditions dans lesquelles on a obtenu des résultats lors de la première expérience. Il y a donc bien un écart significatif entre les résultats des deux expériences.

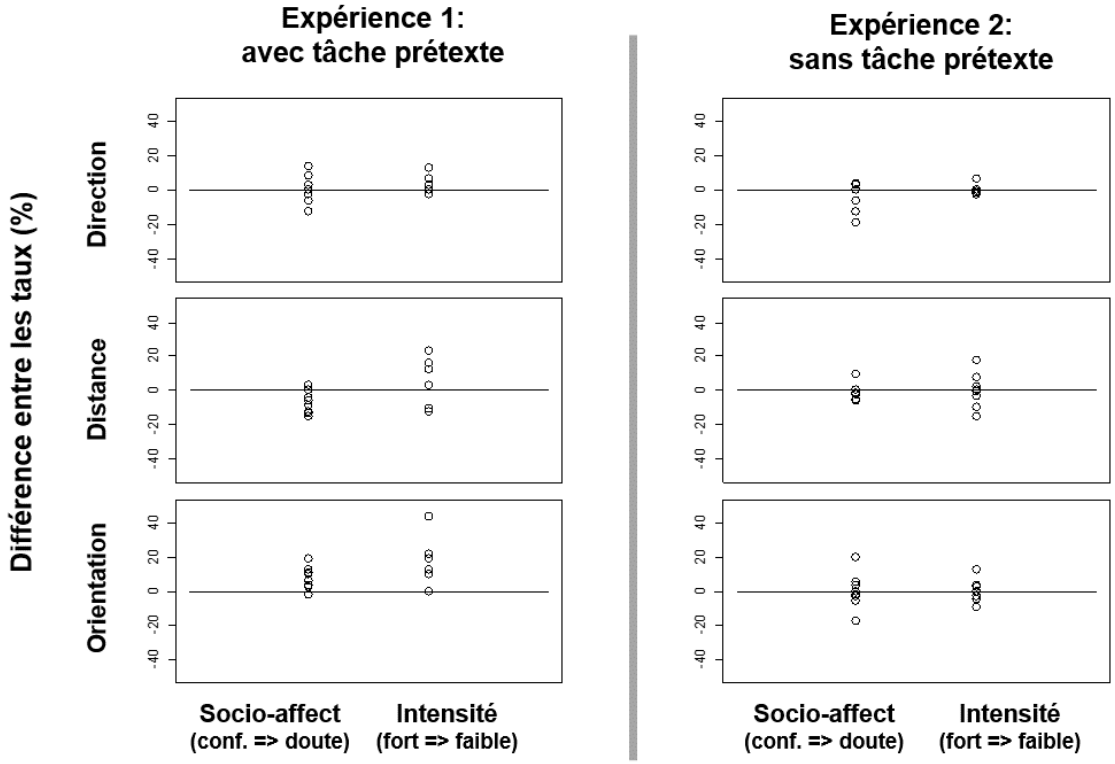


Figure 74 : Écart entre les taux de reconnaissance de chaque catégorie

Tableau 25 : Écart entre les deux taux de reconnaissance de chaque catégorie (moyenne et écart-type)

	Expérience 1		**Expérience 2**	
	Socio-affect (conf. => doute)	**Intensité (fort => faible)**	**Socio-affect (conf. => doute)**	**Intensité (fort => faible)**
direction	- 0,7 ± 2,4 %	+ 2,8 ± 1,4 %	- 3,5 ± 3,0 %	+ 0,5 ± 1,3 %
distance	- 7,8 ± 2,1 %	+ 2,7 ± 4,3 %	- 0,3 ± 2,2 %	- 0,1 ± 3,6 %
orientation	+ 7,4 ± 1,9 %	+ 15,4 ± 3,7 %	- 0,1 ± 3,8 %	+ 0,5 ± 2,3 %

Tableau 26 : Comparaison des résultats des expériences 1 et 2
(valeur-p mesurée à l'aide d'une Anova)

	Socio-affect (conf. => doute)	**Intensité (fort => faible)**
direction	0,46	0,26
distance	0,028 *	0,63
orientation	0,078 .	0,0059 **

4.2.3 Résumé

Une version sans prétexte de l'expérience a été testée. Ses résultats sont significativement différents de ceux de la première expérience : cette fois, on n'observe aucune variation significative des taux de reconnaissance en fonction du type de voix utilisé. Cependant, la variabilité inter-sujet n'a pas pour autant diminué d'une expérience à l'autre : là où les premiers sujets avaient tendance à se tromper tous dans le même sens, ceux de la seconde expérience ont chacun été influencé dans un sens différent. Une hypothèse pour expliquer ce phénomène est que dans le cas où les sujets sont conscients des variations prosodiques, ils essayent d'en tenir compte pour corriger leur estimation de la position spatiale, mais n'utilisent pas tous la même stratégie.

Ces résultats sont donc très intéressants, car ils confirment la nécessité d'utiliser une tâche prétexte dans ce type d'études : lorsque les facteurs sous-jacents apparaissent clairement aux yeux des sujets, leur perception n'est pas la même que lorsque ces facteurs sont cachés. Cela justifie la démarche méthodologique annoncée au chapitre 3.

Néanmoins, ces deux premières expériences souffrent d'une limite importante : le nombre de participants est bien trop faible pour pouvoir tirer une conclusion consistante. De plus, les stimuli utilisés n'ont pu être vérifiés qu'a posteriori : quelques-uns sont inévitablement non conformes aux caractéristiques de productions attendues, ce qui a certainement influencé nos résultats, en accroissant leur variabilité. Pour répondre à ces critiques, une troisième répétition de l'expérience a donc été réalisée, cette fois sous la forme d'un test en ligne.

4.3 Troisième expérience : un test en téléprésence

Les deux expériences précédentes nous ont permis de nous faire une idée de la manière dont le type de voix peut influencer les perceptions spatiales dans le cas d'une écoute directe. Nous avons ensuite souhaité étudier le cas où l'écoute se fait à distance, à travers un robot de téléprésence. Cette fois, les stimuli ont été enregistrés à l'avance, à l'aide de notre robot de téléprésence. L'objectif était double : d'un côté, vérifier que Robair permet à son pilote de repérer la position de ses interlocuteurs dans l'espace ; de l'autre, obtenir des résultats supplémentaires sur la perception de la distance dans des conditions plus contrôlées. On vérifiera également que le type de voix est bien reconnu.

4.3.1 Protocole expérimental

Commençons par présenter les spécificités de cette expérience. Tout d'abord, nous expliqueront la manière dont les stimuli ont été enregistrés ; puis nous décrirons en détails l'interface du test en ligne, et son découpage en plusieurs parties.

4.3.1.1 Stimuli utilisés

Une liste de 8 mots-clés a été sélectionnée parmi 43 utilisés dans les expériences précédentes. Il s'agit des mots : fleur d'oranger, eucalyptus, champignon, ananas, brûlé, coco, pin et rose. Ils ont été sélectionnés de manière à faire varier le nombre de syllabes et le contenu phonétique. En effet, on veut vérifier si la perception est robuste, ou si elle dépend au contraire du mot-clé

choisi : on peut imaginer par exemple que plus la durée du mot-clé est longue, plus les sujets ont le temps pour bien le localiser.

Ces 8 mots-clés ont été enregistrés dans les nouveaux locaux de Domus. Le robot Robair était placé dans un coin du salon, tandis que je le téléopérais depuis la chambre sourde (cf. Figure 75). Dans la pièce du robot, Aubergé prononçait la liste de mots-clés pour chaque classe de socio-affect × intensité, puis se déplaçait à la position suivante. Si elle n'était pas satisfaite par une de ses productions, elle pouvait la répéter autant de fois qu'elle le souhaitait. Cette méthode nous a permis de recueillir un corpus bien plus proche de notre idée initiale, avec une opposition entre une voix polie / sympathique, et une voix autoritaire / distante.

La voix d'Aubergé était enregistrée par le robot, et transmise à mon ordinateur, sur lequel était lancé un enregistrement de monitoring, afin de conserver le signal diffusé dans mon casque audio. Elle portait également un micro serre-tête, placé près de sa bouche. Comme la salle n'était pas assez grande pour pouvoir tourner autour du robot, nous le faisions pivoter de 90° entre chaque changement de direction. Chaque mot-clé existe donc en deux versions : une enregistrée avec le robot de téléprésence, et l'autre enregistrée simultanément au niveau de la bouche de la locutrice. Les enregistrements entendus par les participants du test sont ceux enregistrés avec le robot ; les autres constituent des signaux de référence, d'excellente qualité.

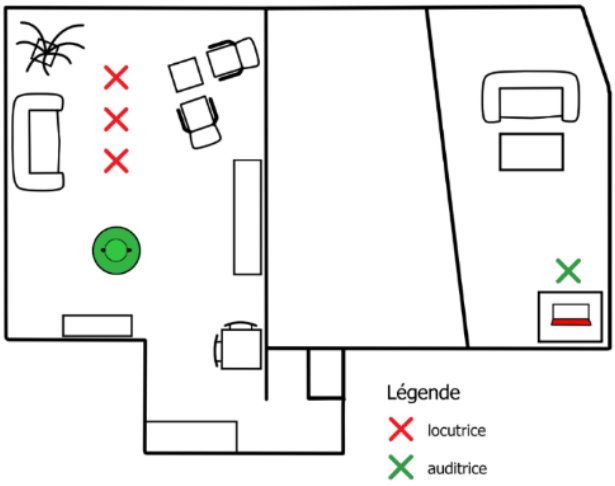


Figure 75 : Vue du dessus du dispositif d'enregistrement

Tout d'abord, les enregistrements ont été grossièrement découpés pour isoler les mots clés. Aubergé les a annotés à l'aide d'une interface web, lui permettant d'écouter chaque enregistrement de référence l'un après l'autre, et de cocher la ou les qualités de voix correspondantes. Les sons dont la qualité de voix n'était pas satisfaisante ont été supprimés. Ainsi, tous les sons étiquetés comme *harsh* ou *creaky* ont été éliminés. Tous les « doutes » conservés sont *breathy,* et toutes les « confiances » sont *tense*.

Enfin, la durée des enregistrements a été standardisée automatiquement, de manière à avoir 200 ms de silence avant et après chaque enregistrement.

Notons qu'il existe un bruit basse-fréquence, dû au système de ventilation de l'appartement, et audible sur les enregistrements du robot.

4.3.1.2 Présentation de l'expérience en ligne

L'expérience se déroulait en ligne, via un site web hébergé sur les serveurs de l'Imag, l'institut d'Informatique et de Mathématiques Appliquées de Grenoble. Pour concevoir l'interface de test, nous nous sommes inspirés du travail effectué par Clarisse Bayol pendant son stage de Master 2 (2016). Tout d'abord, les participants arrivaient sur une page de présentation de l'expérience (cf. Figure 76) : l'objectif annoncé était d'étudier la qualité de l'immersion acoustique de notre robot de téléprésence. Suivait une page d'information, invitant les sujets à se munir d'un casque ou d'une paire d'écouteurs, et à régler leur volume à l'aide de deux exemples, correspondant aux sons les plus faibles et les plus forts du test. Les participants arrivaient ensuite sur une page de renseignement, où ils devaient indiquer leur genre, âge, langue maternelle, s'ils avaient une expérience particulière en acoustique, ou des problèmes d'audition, s'ils avaient l'habitude d'écouter des enregistrements au casque, et si oui dans quel contexte (musique, film, jeux vidéo, appels audio ou vidéo, ou autre). On leur demandait également de juger a priori de la qualité de leur casque / paire d'écouteur. Les participants arrivaient alors au test proprement dit, précédé d'une page de consignes expliquant la tâche demandée.

Afin de faciliter l'analyse des résultats, et simplifier la tâche demandée aux sujets, l'expérience a été découpée en trois tests courts d'environ 5 min chacun : 1) perception de la distance, 2) perception du socio-affect, 3) perception de la direction. À la fin de chaque petit test, les sujets pouvaient laisser un commentaire, et il leur était proposé de passer au test suivant. S'ils acceptaient, ils étaient dirigés vers une nouvelle page de consignes, puis vers le test. L'ordre de présentation des trois tests a été choisi de manière à privilégier les résultats les plus intéressants : ceux sur la distance. Nous espérions qu'ainsi les sujets seraient le plus naïfs possibles au moment d'estimer la distance : même s'ils percevaient les variations prosodiques, ils n'étaient mis au courant des deux catégories étudiées qu'à partir du test 2. Enfin, le test 3 était prévu pour vérifier que le système d'écoute binaurale du robot permet bien au pilote de repérer grossièrement la direction d'arrivée des sons.

En réalité, le découpage du test était un peu plus complexe que tel que présenté aux participants : le test sur la distance était lui-même subdivisé en 4 tests distincts. En effet, il semblait peu pertinent de demander aux sujets une estimation directe de la distance, sachant que cette tâche est déjà très difficile en présentiel. Nous avons donc opté pour une présentation des stimuli deux par deux : les sujets devaient ainsi donner uniquement une estimation relative de la distance, en indiquant si à leur avis, entre les deux sons, la locutrice n'avait pas bougé, s'était rapprochée, ou s'était éloignée. Or, en présentant les stimuli par pairs, on multiplie drastiquement le nombre de combinaisons possibles : il a donc fallu subdiviser le test, en sélectionnant chaque fois soigneusement les paramètres étudiés, et leurs plages de variations. Ces choix sont récapitulés dans les tableaux 27, 28 et 29.

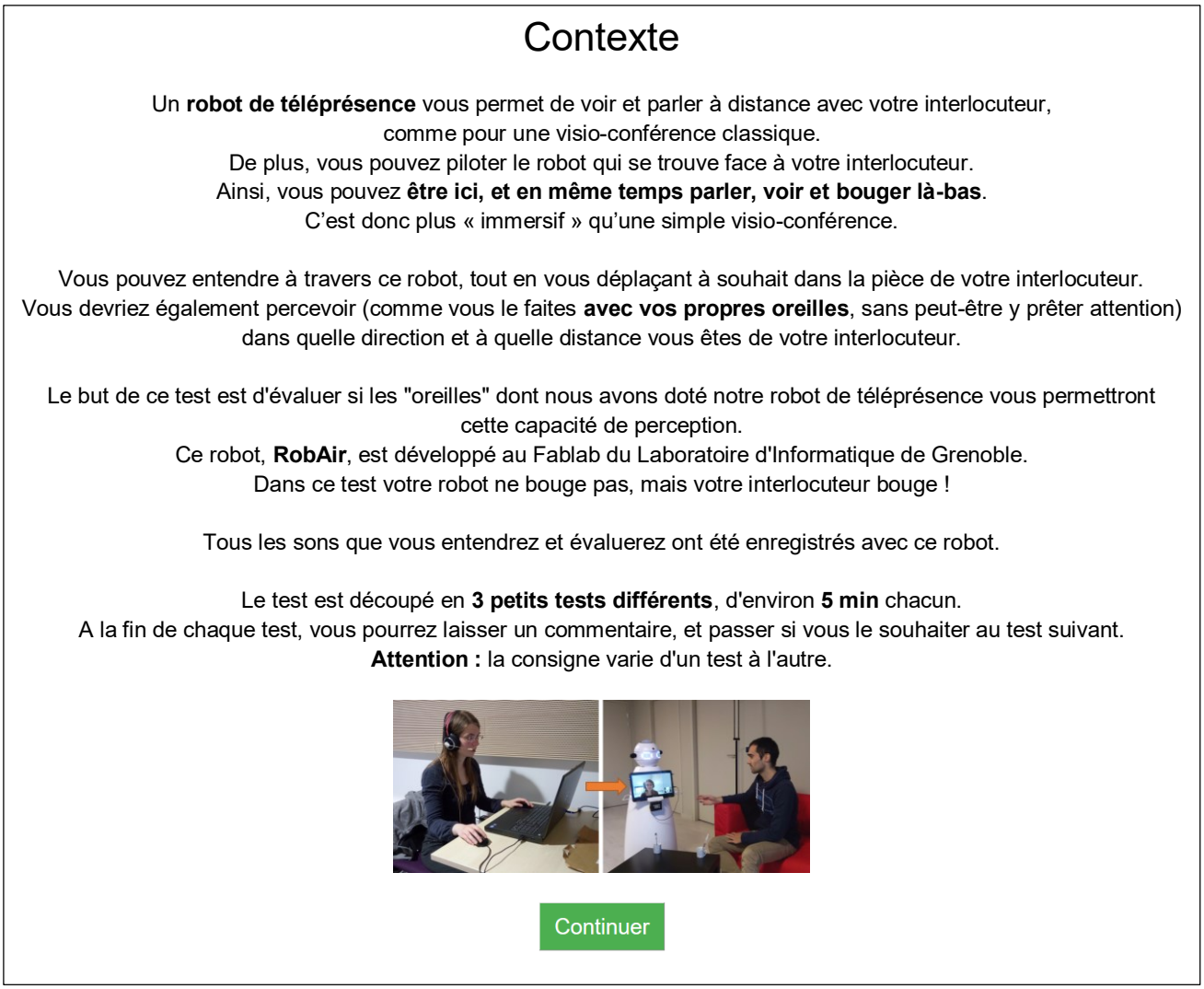


Figure 76 : Page d'accueil du test en ligne

Pour les tests sur la **distance**, seuls les mots de longueurs moyennes (champignon, ananas, brûlé, coco) sont utilisés. Pour chaque nouvelle paire de sons, le mot suivant de la liste est sélectionné. Une fois que la liste entière de mots a été parcourue, elle est mélangée par permutation circulaire. La sélection de la direction se fait de la même manière. Ces deux paramètres sont donc des paramètres « bruits » : ils varient pour que les sujets n'entendent pas toujours la même chose, et continuent à croire au prétexte ; mais leur influence sur les résultats n'est pas directement notre objet d'étude. Les variables qui nous intéressent ici sont la distance, l'orientation, le socio-affect et l'intensité.

Pour le test 1), la distance est fixée, ce qui signifie que les deux sons de chaque paire ont été enregistrés à la même distance. En revanche, le socio-affect et l'intensité varie à l'intérieur de chaque paire : les deux sons ne peuvent pas appartenir tous les deux à la même classe (« doute faible », « doute fort », « confiance faible » ou « confiance forte »). Pour ce premier test, il y a donc 36 paires de sons :

3 (distances) × 4 (socio-affect et intensité du son1)
× 3 (socio-affect et intensité du son2 différente de celle du son1).

Pour le test 2, c'est le socio-affect et l'intensité qui sont fixés, et la distance, ou l'orientation, qui varie. Le nombre de paires est donc de 32 :

4 (socio-affect et intensité) × 3 (distance du son1) × 2 (distance du son2)
\+ 4 (socio-affect et intensité) × 2 (orientation du son1) × 1 (orientation du son2)

Pour les tests 3 et 4, les deux sons de chaque paire doivent être de distance, socio-affect et intensité différente. Le nombre de paires total s'élève donc à 72 :

3 (distance du son1) × 2 (distance du son2) × 4 (socio-affect et intensité du son1)
× 3 (socio-affect et intensité du son2)

Ces paires ont été séparées en deux groupes arbitraires X et O, utilisés respectivement pour les tests 3 et 4. La règle de découpage des groupes est représentée dans le Tableau 28 : elle permet de faire varier l'ordre des distances d'enregistrements des sons 1 et 2.

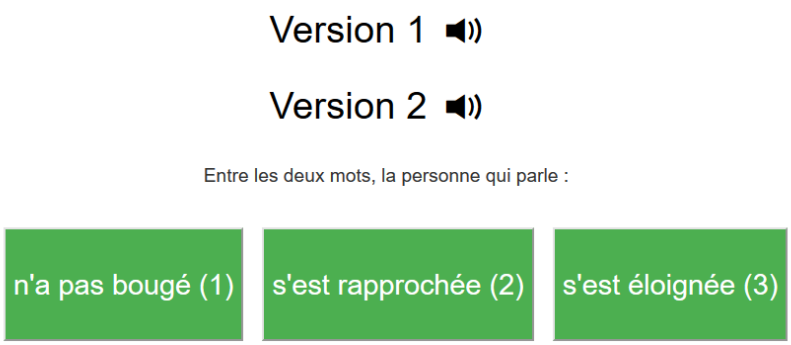


Figure 77 : Capture d'écran de l'interface pour les tests 1-4 sur la distance

En fonction de son numéro d'arrivée, chaque sujet se voyait attribuer un des quatre tests sur la distance. Ensuite, il passait au test 5, sur le **socio-affect**. Pour ce test, seul un mot était joué à la fois, et le sujet devait décider si la personne enregistrée était plutôt polie / sympathique, ou autoritaire. Nous avons préféré ces deux étiquettes à celles utilisées précédemment (doute / confiance), car elles sont plus proches des socio-affects recherchés initialement, mais qui n'avaient pas pu être produits parfaitement dans la première expérience pour pouvoir maintenir le faux-prétexte.

Cette fois, la direction d'arrivée est à nouveau considérée comme un paramètre « bruit ». Les mots utilisés sont uniquement les plus courts (pin, rose) et les plus longs (eucalyptus, fleur d'oranger). Les paramètres variables sont le nombre de syllabes du mot-clé (1 ou 4), la distance, l'orientation, l'intensité et le socio-affect. Ainsi, 32 combinaisons différentes sont testées :

2 (nombre de syllabes) × 4 (distance et orientation) × 4 (socio-affect et intensité du son)

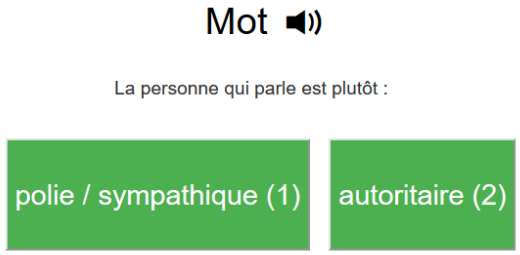


Figure 78 : Capture d'écran de l'interface pour le test 5 sur le socio-affect

Enfin, le dernier test concerne la perception de la **direction** d'arrivée. Cette fois, les 64 combinaisons des expériences 1 et 2 étaient utilisées :

4 (direction) × 4 (distance et orientation) × 4 (socio-affect et intensité du son)

Le mot-clé était choisi en suivant la liste complète des huit mots, mélangés par permutation circulaire dès que la fin de la liste était atteinte.

Les temps de réaction des sujets sont également enregistrés : il s'agit de la durée entre le moment où un nouveau son est joué, et où la personne clique sur un bouton.

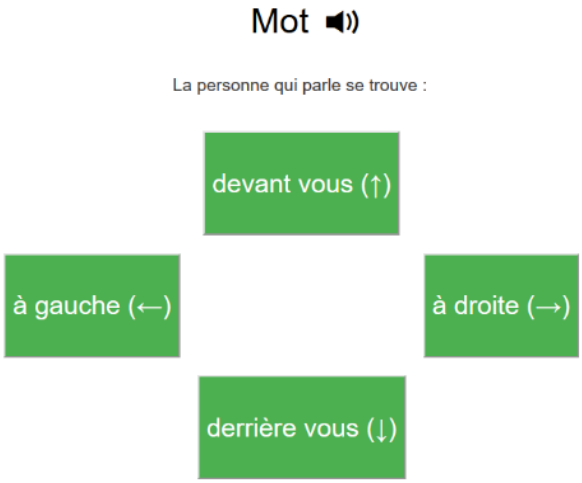


Figure 79 : Capture d'écran de l'interface pour le test 6 sur la direction

Question posée aux sujets	N°	Présentation des stimuli	Paramètre fixé (cardinal F)	Paramètre variable (cardinal V)	Décompte si paire de sons : F × V × (V – 1) si son unique : F × V
Distance	1)	2 par 2	distance (3)	socio-affect + intensité (4)	3 × 4 × 3 = 36 paires
	2)	2 par 2	socio-affect + intensité (4)	distance (3) orientation (2)	(4 × 3 × 2) + (4 × 2 × 1) = 32 paires
	3)	2 par 2	∅	distance (3) socio-affect + intensité (4)	(3 × 2 × 4 × 3) / 2 = 36 paires - groupe X
	4)	2 par 2	∅	distance (3) socio-affect + intensité (4)	(3 × 2 × 4 × 3) / 2 = 36 paires - groupe O
Socio-affect	5)	1 par 1	∅	nombre de syllabes (2) distance + orientation (4) socio-affect + intensité (4)	2 × 4 × 4 = 32 uniques
Direction	6)	1 par 1	∅	tous (4 × 4 × 4)	4 × 4 × 4 = 64 uniques

Tableau 27 : Liste des tests

Question posée aux sujets	N°	Nombre de syllabes	Mots utilisés	Paramètre « bruit »
Distance	1)	2 ou 3	coco, brûlé, ananas, champignon	direction (4) mot (4)
	2)			
	3)			
	4)			
Socio-affect	5)	1 ou 4	pin, rose, eucalyptus, fleur d'oranger	direction (4) mot (4)
Direction	6)	1, 2, 3 ou 4	tous	mot (8)

Tableau 29 : Liste des mots utilisés pour chaque test

son1 / son2	proche	milieu (face)	loin
proche		X	O
milieu (face)	O		X
loin	X	O	

Tableau 28 : Définition des groupes pour les tests 3 et 4

4.3.2 Caractéristiques des enregistrements

Avant de passer à l'analyse des résultats, nous allons étudier les enregistrements, afin de mettre en évidence les indices acoustiques susceptibles d'être utilisés pour estimer la distance et la direction. Nous nous intéresserons en particulier à l'intensité moyenne mesurée en fonction des différentes variables. Les deux enregistrements seront utilisés : celui pris au niveau de la bouche, et celui du robot de téléprésence. Pour calculer les intensités moyennes, seules les intensités supérieures à 35 dB sont prises en compte.

Comme prévu, l'intensité varie significativement en fonction du type de voix utilisée (cf. Figure 80). Il faut donc en tenir compte pendant l'analyse des résultats si l'on souhaite mettre en évidence des variations plus subtiles. Les modèles linéaires mixtes sont un moyen d'y parvenir. Par exemple, pour étudier la variation d'intensité en fonction de la distance, on peut calculer les coefficients du modèle suivant :

$$intensité\ moyenne \sim distance + (1 \mid type\ de\ voix)$$

Pour plus de détails sur ces modèles, le lecteur est renvoyé à l'Annexe C, inspirée du tutoriel de (Winter 2013).

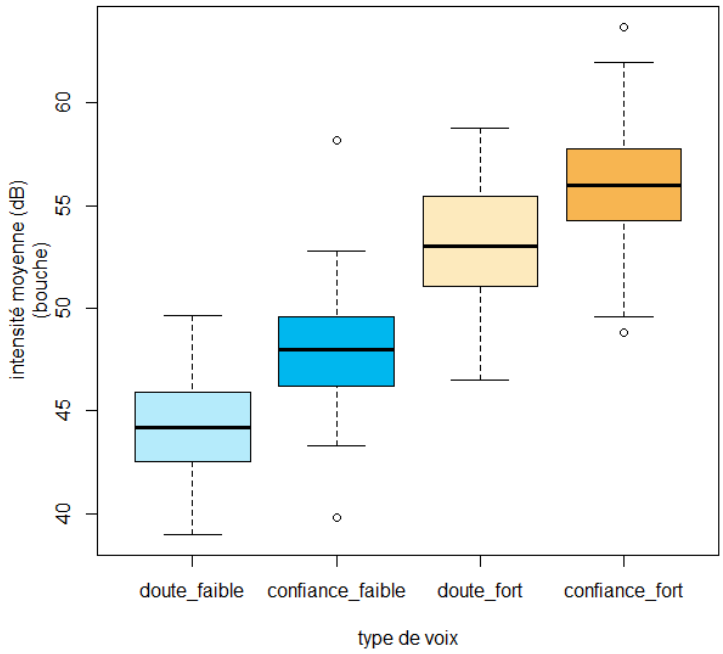


Figure 80 : Intensité moyenne mesurée en fonction du type de voix à partir des enregistrements pris au niveau de la bouche

Le Tableau 30 représente les résultats de la modélisation de la distance et de l'orientation. La première ligne correspond à la valeur moyenne estimée pour les sons de la catégorie « proche face », qui sert de référence. Les autres lignes du tableau contiennent l'écart estimé entre la catégorie étudiée et la catégorie de référence : ce sont les pentes du modèle linéaire.

On constate que pour les sons enregistrés au niveau de la bouche, l'intensité moyenne ne varie pas en fonction de la distance ou de l'orientation, puisque l'erreur-type sur la pente est supérieure à la valeur estimée. On en déduit que la force de voix de la locutrice est relativement stable, quelle que soit la distance du robot.

En revanche, pour les sons enregistrés par le robot, l'intensité moyenne diminue effectivement avec la distance, puisque la pente entre les distances « proche » et « milieu » est d'environ - 0,7 ± 0,3 dB, et celle entre les distances « proche » et « loin » d'environ - 1,3 ± 0,3 dB. On note également que la diminution d'intensité est plus importante pour les sons de la catégorie « milieu dos », que la catégorie « milieu face ».

Tableau 30 : Coefficients du modèle linéaire mixte modélisant l'intensité moyenne en fonction de la distance et de l'orientation de la locutrice

Distance + orientation	Intensité moyenne (bouche)	Intensité moyenne (robot)
proche face (intercept)	50,4 ± 2,3 dB	49,6 ± 2,0 dB
milieu face	- 0,2 ± 0,4 dB	- 0,7 ± 0,3 dB
milieu dos	- 0,5 ± 0,4 dB	- 2,1 ± 0,3 dB
loin face	+ 0,1 ± 0,4 dB	- 1,3 ± 0,3 dB

Le Tableau 31 rapporte les résultats de la modélisation de l'intensité en fonction de la direction. On constate à nouveau que l'intensité est stable au niveau de la bouche. En revanche, l'intensité moyenne enregistrée par le robot est légèrement plus élevée pour les sons enregistrés à gauche, et à droite. En effet, dans ce cas, un des microphones est dirigé directement vers la source de son, et capte donc mieux les ondes sonores qu'en étant à un angle de 90°. Par ailleurs, la dernière colonne du tableau confirme que l'intensité du canal gauche est plus forte que celle du canal droit lorsque le son vient de la gauche, et inversement.

On note également que lorsque le son vient de l'avant, l'intensité du canal gauche est légèrement plus forte que celle du canal droit ; et inversement lorsque le son vient de derrière. Ceci est probablement dû aux propriétés acoustiques de la pièce et à la manière dont les ondes sonores s'y réverbèrent : le micro qui capte le plus d'intensité est tourné vers un angle de la pièce, tandis que l'autre est tourné vers la cuisine, un espace plus dégagé.

Tableau 31 : Coefficients du modèle linéaire mixte modélisant l'intensité moyenne en fonction de la direction de la locutrice

Direction	Intensité moyenne (bouche)	Intensité moyenne (robot)	Différence d'intensité entre les deux canaux du robot (L – R)
avant (intercept)	50,4 ± 2,3 dB	48,0 ± 2,0 dB	0,7 ± 0,1 dB
gauche	+ 0,1 ± 0,4 dB	+ 0,9 ± 0,3 dB	+ 0,5 ± 0,1 dB
droite	- 0,1 ± 0,4 dB	+ 1,2 ± 0,3 dB	- 1,4 ± 0,1 dB
derrière	- 0,5 ± 0,4 dB	- 0.2 ± 0,3 dB	- 1,0 ± 0,1 dB

Le timbre des enregistrements varie également en fonction de la distance, car plus le robot est éloigné, plus le rapport d'intensité entre la voix directe et ses multiples réverbérations dans la pièce diminue. À titre de comparaison, la Figure 81 et la Figure 82 représentent le spectre moyen des enregistrements, obtenu par tiers d'octaves entre 50 et 8 kHz, et en regroupant les sons par catégorie de distance et d'orientation. Dans le cas de la Figure 81, il s'agit des enregistrements effectués au niveau de la bouche : les quatre courbes sont indissociables les unes des autres. Au contraire, la Figure 82 représente les spectres moyens des enregistrements du robot, obtenus en moyennant les spectres des deux canaux audio. Cette fois, on constate des variations en fonction de la distance et de l'orientation ; en particulier dans les hautes fréquences, il est possible de classer les quatre groupes de sons par énergie décroissante : « proche » > « milieu_face » > « loin » > « milieu_dos ». Ainsi l'orientation de la locutrice a plus d'impact sur son timbre que la distance seule. Notons que ces variations sont relativement faibles : l'écart maximal est d'environ 5dB entre la courbe « proche » et la courbe « milieu_dos » pour la bande de fréquence 3150 – 5000 Hz. Les variations liées au type de voix sont bien plus importantes : par exemple, l'écart entre le « doute faible » et la « confiance forte » dépasse les 15 dB dans la bande 400 – 6000 Hz (cf. Figure 82). Notons que le maximum d'énergie dépend également de l'intensité et du socio-affect : situé autour de 300 Hz pour le doute faible, il se décale vers les hautes fréquences pour la confiance, et lorsque l'intensité augmente (cf. Tableau 32).

Tableau 32 : Coefficients du modèle linéaire mixte modélisant l'indice fréquentiel du maximum d'énergie en fonction du type de voix utilisé

Type de voix	Fréquence centrale du maximum d'énergie
doute faible	293 ± 38 Hz
confiance faible	+ 141 ± 19 Hz
doute fort	+ 170 ± 19 Hz
confiance forte	+ 261 ± 17 Hz

En conclusion, nous avons pu observer des variations d'intensité permettant de deviner la provenance du son. Ainsi, comme prévu, l'intensité sonore diminue avec la distance d'enregistrement, et on observe des différences d'intensité entre les deux « oreilles » du robot, ainsi que des variations de timbre. Cependant, ces variations sont très subtiles : seulement 1 à 2 dB. Il reste à savoir si elles sont suffisantes pour être perçues par les sujets, et s'ils parviennent à les distinguer efficacement des variations prosodiques.

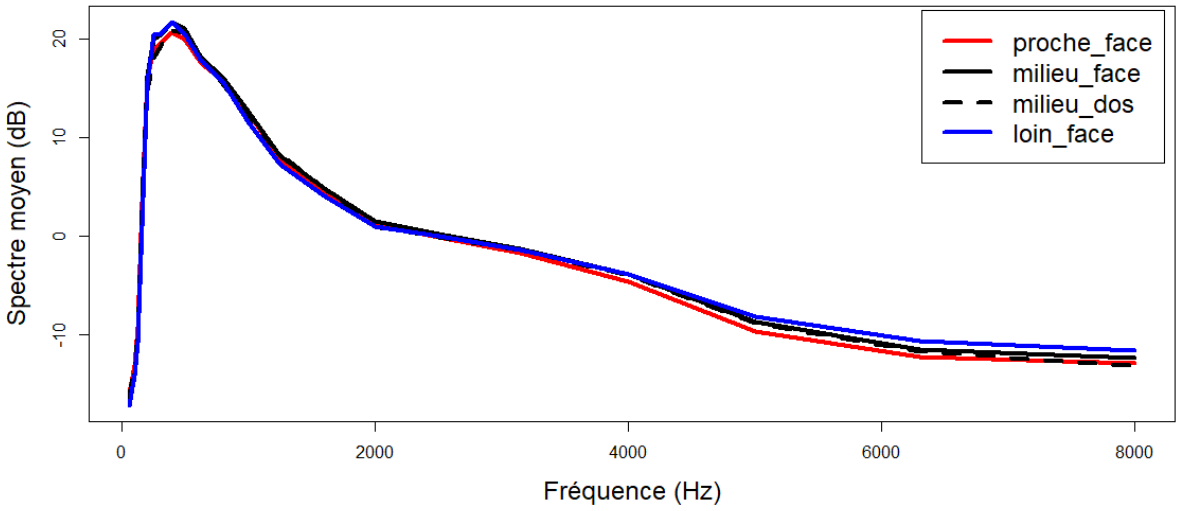


*Figure 81 : Spectre moyen en fonction de la distance (sons enregistrés au niveau de la **bouche**)*

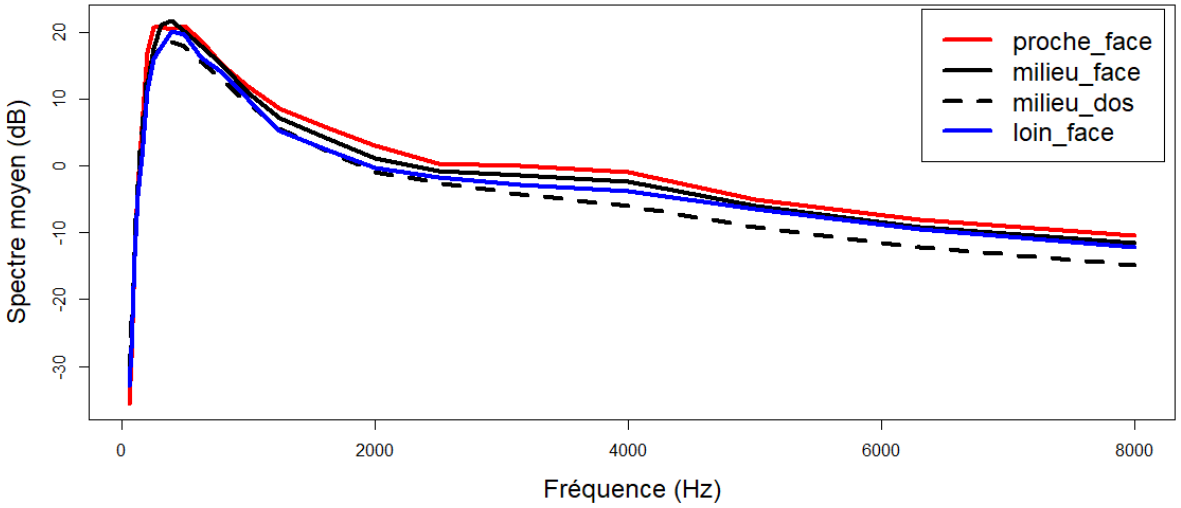


*Figure 82 : Spectre moyen en fonction de la distance (sons enregistrés par le **robot**)*

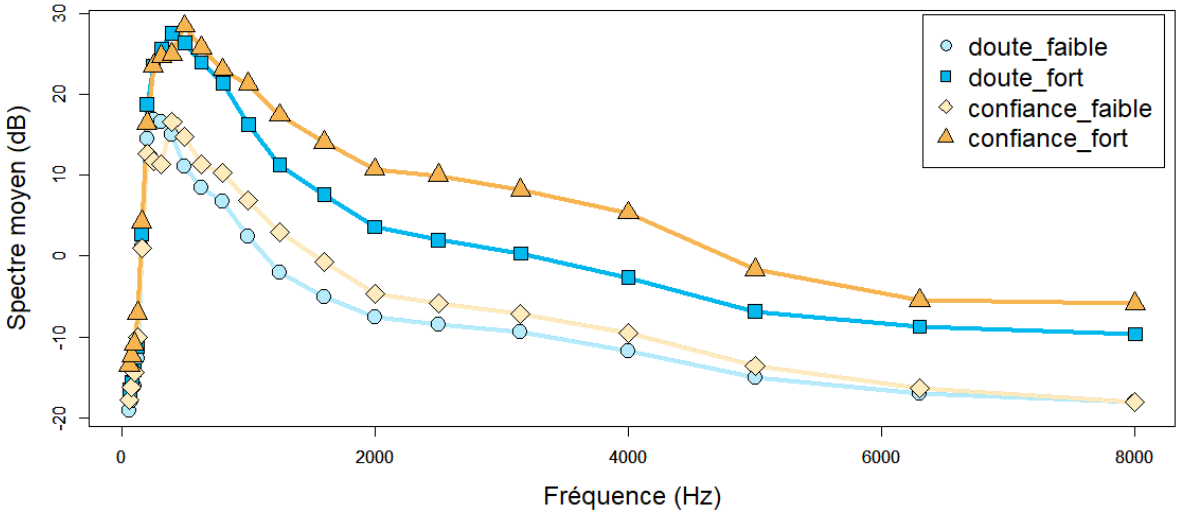


Figure 83 : Spectre moyen en fonction du type de voix (sons enregistrés au niveau de la bouche)

4.3.3 Analyse statistique des résultats

Dans cette partie, nous présenterons les résultats des tests en ligne. Nous commencerons par décrire la cohorte des participants, puis nous étudierons chaque test, l'un après l'autre.

4.3.3.1 Participants

65 personnes se sont connectées au site, mais 16 se sont arrêtées avant d'atteindre la fin du premier test, et ont été retirées des résultats. Sur les 49 personnes restantes, on compte 22 femmes et 26 hommes. Leur âge va de 21 à 69 ans. La moitié d'entre elles avait moins de 30 ans (cf. Figure 84). À trois exceptions près, toutes avaient comme langue maternelle le français. Une seule personne a indiqué avoir des problèmes d'audition. Un quart des participants se considèrent comme des experts en acoustique, du fait de leur activité professionnelle ou de leurs loisirs. Les deux tiers indiquent avoir l'habitude d'écouter au casque ou avec des écouteurs, principalement pour écouter de la musique, regarder des films, ou effectuer des appels audio ou vidéo.

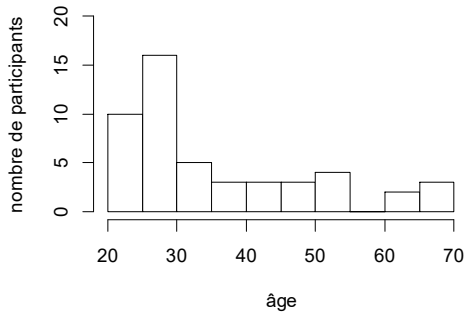


Figure 84 : Histogramme de l'âge des participants

Le nombre de participants diminue entre chaque partie du test : ainsi, sur les 49 participants initiaux, qui ont complétés un des tests 1-4 sur la distance, on ne compte plus que 28 participants ayant complété entièrement le test 5 sur le socio-affect, et 32 pour le test 6 sur la direction[25]. Le détail du nombre de participants pour les tests sur la distance est fourni dans le Tableau 33.

Tableau 33 : Détail du nombre de participants pour les tests sur la distance

	test 1	test 2	test 3	test 4
Nombre de participants	15	11	11	13

[25] Étant donné la manière dont le test a été conçu, on devrait avoir plus de sujets au test 5 qu'au test 6. Cependant, une erreur dans l'interface du test 5 a conduit certains sujets à ne pas compléter entièrement le test, et à passer directement au test 6.

4.3.3.2 Perception de la distance (tests 1 à 4)

Nous commencerons par étudier les résultats du test sur la distance. Rappelons qu'il existe quatre versions différentes de ce test : pour le test 1, la distance des sons de chaque couple est fixée, tandis que le type de voix varie ; pour le test 2, c'est le type de voix qui est fixé, et la distance qui varie entre les deux sons ; enfin, les tests 3 et 4 font entendre toutes les combinaisons restantes.

Il est possible d'étudier chaque test séparément, de manière à faire ressortir les résultats spécifiques de chaque test. Par soucis de concision pour le lecteur, nous avons préféré regrouper les résultats des quatre tests en un unique jeu de données. On utilisera les notations suivantes : (1 > 2) lorsque le son 2 est perçu plus proche que le son 1, (1 = 2) lorsque les deux sons sont perçus à la même distance, et (1 < 2) lorsque le son 2 est perçu plus loin que le son 1.

La matrice de confusion globale est représentée dans le Tableau 34. On constate que le taux de reconnaissance est assez faible : seulement 50 % en moyenne ; ce qui signifie que les sujets n'ont pas perçu la bonne distance une fois sur deux. On note également que les sujets perçoivent plus souvent un rapprochement entre les deux sons, plutôt qu'un éloignement ((1 > 2) : 31 % | (1 = 2) : 45 % | (1 < 2) : 23 %). Plus précisément, sur les 49 participants, 38 ont cliqué plus souvent sur le bouton « rapprochement » que sur le bouton « éloignement ». En conséquence, le taux de reconnaissance est plus élevé lorsque le deuxième son est proche (54 % contre 42 %). Pourtant le jeu de données est relativement équilibré. Pour expliquer ce résultat, nous proposons deux interprétations : soit les sujets percevaient plus facilement un rapprochement entre les deux sons qu'un éloignement, soit ils ont été influencés par l'ordre des boutons dans l'interface (de gauche à droite : « même distance », « rapprochement », puis « éloignement).

Tableau 34 : Matrice de confusion pour la perception de la distance

Distance		**Production**		
		(1 > 2)	(1 = 2)	(1 < 2)
Perception	(1 > 2)	299 **54 %**	165 27 %	75 14 %
	(1 = 2)	210 38 %	326 **52 %**	241 44 %
	(1 < 2)	41 7 %	129 21 %	226 **42 %**

Nous avons choisi de représenter les résultats sous une forme matricielle. Pour commencer, la Figure 85 représente la répartition des réponses des sujets en fonction de l'**intensité**. Les colonnes correspondent aux étiquettes du premier son du couple, et les lignes aux étiquettes du deuxième son. La matrice violette, en haut à gauche, correspond au décompte des couples perçus à une distance fixe (1 = 2). La matrice marron, en haut à droite, correspond au décompte

des couples pour lequel le deuxième son est perçu plus proche que le premier (1 > 2). La matrice verte, en bas à gauche, correspond au décompte des couples pour lequel le deuxième son est perçu plus loin que le premier (1 < 2). Enfin, la matrice verte et marron, en bas à droite, représente la différence entre les deux matrices précédentes, afin de faire ressortir les grandes tendances.

On constate sans surprise que les sujets ont été fortement influencés par l'intensité des sons. Ainsi, pour les couples d'intensité différentes, en moyenne, le son d'intensité forte a été perçu plus proche dans 47 % des cas, à la même distance que le son d'intensité faible dans 41 % des cas, et plus loin dans seulement 12 % des cas.

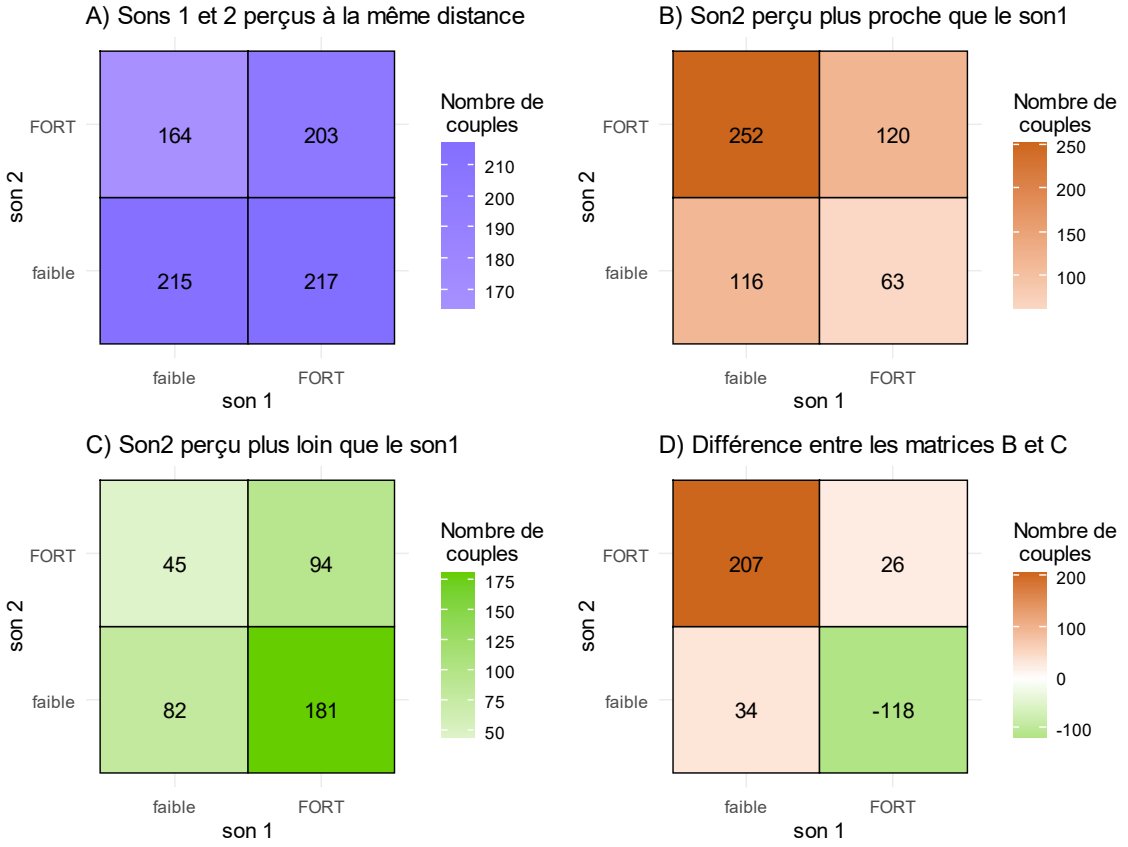


Figure 85 : Perception de la distance en fonction de l'intensité

Pour aller plus loin, la Figure 86 représente la répartition des réponses des sujets en fonction du **type de voix**. Ici, seul le socio-affect est indiqué ; l'intensité est représentée par la casse utilisée (fort = majuscules / faible = minuscules). On retrouve la répartition précédente, mais en plus détaillée ; de manière à pouvoir observer l'effet du socio-affect. Ainsi, si on compte le nombre de couple de socio-affects mixtes, on constate que la confiance est perçue plus proche dans 246 cas, tandis que le doute est perçu plus proche dans 207 cas. A priori, la répartition devrait être d'environ (½, ½). Pour savoir si ce résultat est significatif, on utilise un test du khi-deux pour comparer les deux distributions : on obtient une valeur-p de 0,07. À l'échelle de l'ensemble des résultats, le socio-affect n'a donc pas d'effet significatif. En revanche, l'effet est significatif si

on tient compte uniquement des résultats du test 1. Les valeurs-p de chaque test sont disponibles dans le Tableau 35.

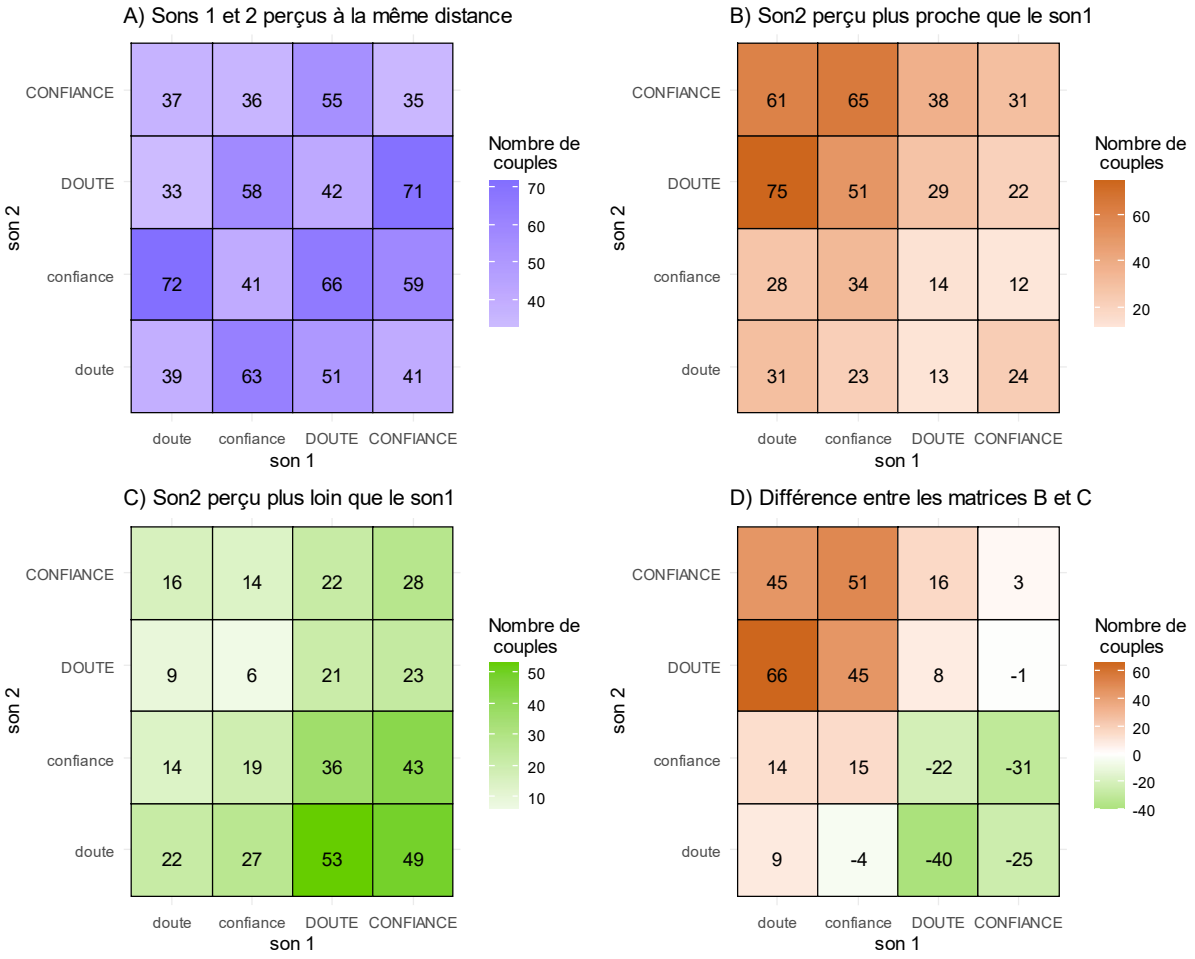


Figure 86 : Perception de la distance en fonction du type de voix

Tableau 35 : Détail des tests statistiques concernant l'effet du type de voix sur la perception de la distance

Facteur étudié	Test	Tendance	Valeur-p (khi-2)	Significativité
intensité	1	faible > fort	$< 2{,}2 \cdot 10^{-16}$	***
	2	NA	NA	NA
	3	faible > fort	$5 \cdot 10^{-8}$	***
	4	faible > fort	$2 \cdot 10^{-14}$	***
socio-affect	1	doute > confiance	$7 \cdot 10^{-4}$	***
	2	NA	NA	NA
	3	doute < confiance	0,4	n.s.
	4	doute > confiance	0,5	n.s.

Passons à présent à l'étude de l'effet des variables spatiales sur la perception de la distance. Tout d'abord, la Figure 87 représente la répartition des couples de sons en fonction de leur **distance** d'enregistrement. On constate que lorsque la distance varie effectivement entre les deux sons du couple, les sujets perçoivent la variation correctement dans 48 % des cas, répondent que la distance n'a pas varié dans 41 % des cas, et se trompent dans le sens de la variation dans seulement 11 % des cas. De plus, le taux de reconnaissance est plus élevé lorsque l'écart de distance est important : en considérant uniquement les couples « proche-loin » et « loin-proche », le taux de bonne reconnaissance monte à 61 %, et le taux d'erreur tombe à 8 %. Pour expliquer ces résultats, on peut supposer qu'en cas d'hésitation, les participants répondent par défaut que la distance est fixe, et qu'ils ne prennent le risque d'opter pour une autre réponse que s'ils sont certains de ce qu'ils ont entendu ; donc lorsque les indices acoustiques permettant d'estimer la distance d'enregistrement sont plus marqués.

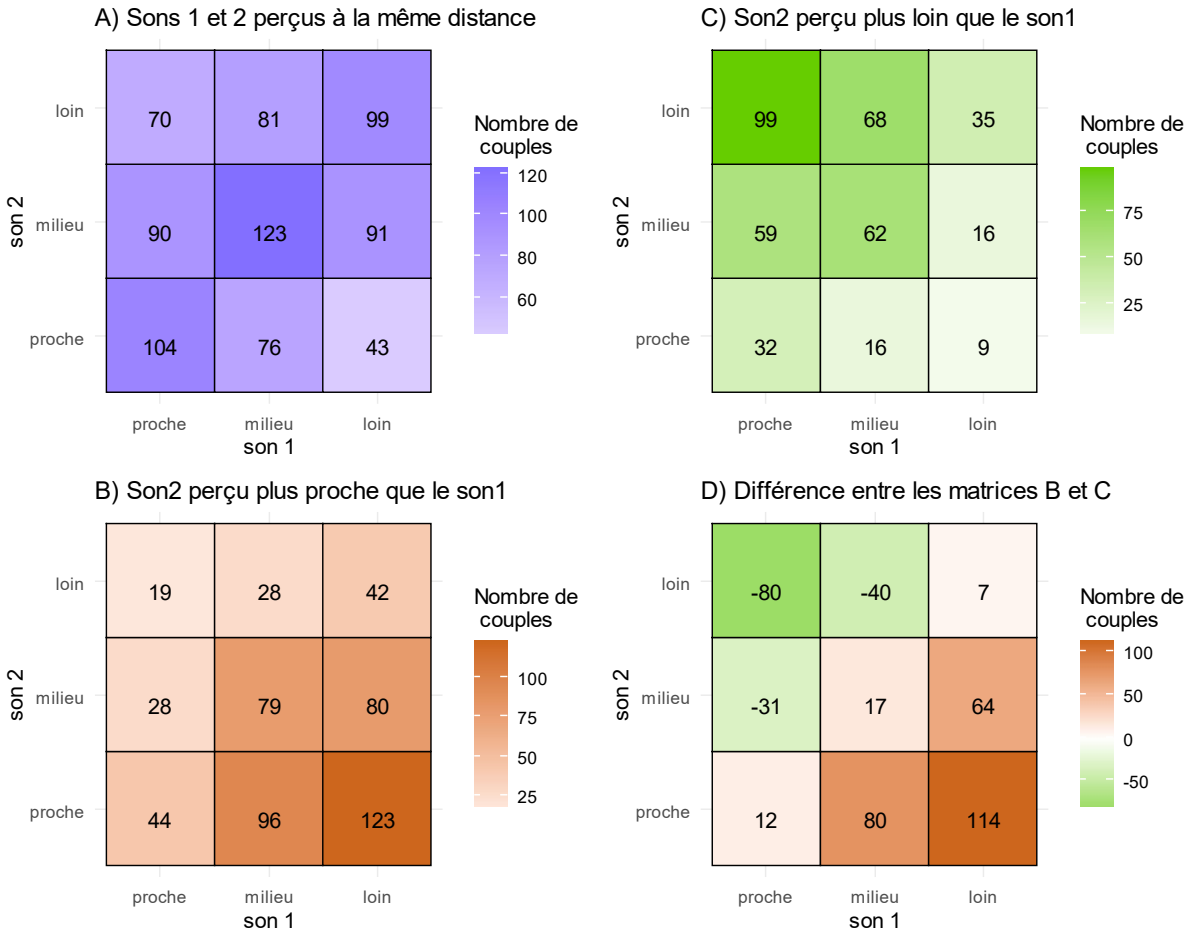


Figure 87 : Perception de la distance en fonction de la distance

La Figure 88 représente la répartition des réponses des sujets en fonction de l'orientation de la locutrice. On constate que dans les couples d'orientation mixte, les sons enregistrés de dos ont été quasi systématiquement perçus plus loin que les sons enregistrés de face ((1 < 2) : 68% | (1 = 2) : 28 % | (1 > 2) : 3 %). Or, il s'agit uniquement de couples du test 2, pour lesquels l'intensité et le socio-affect étaient fixés. La diminution d'intensité due au changement d'orientation a donc été majoritairement interprétée par les sujets comme un éloignement.

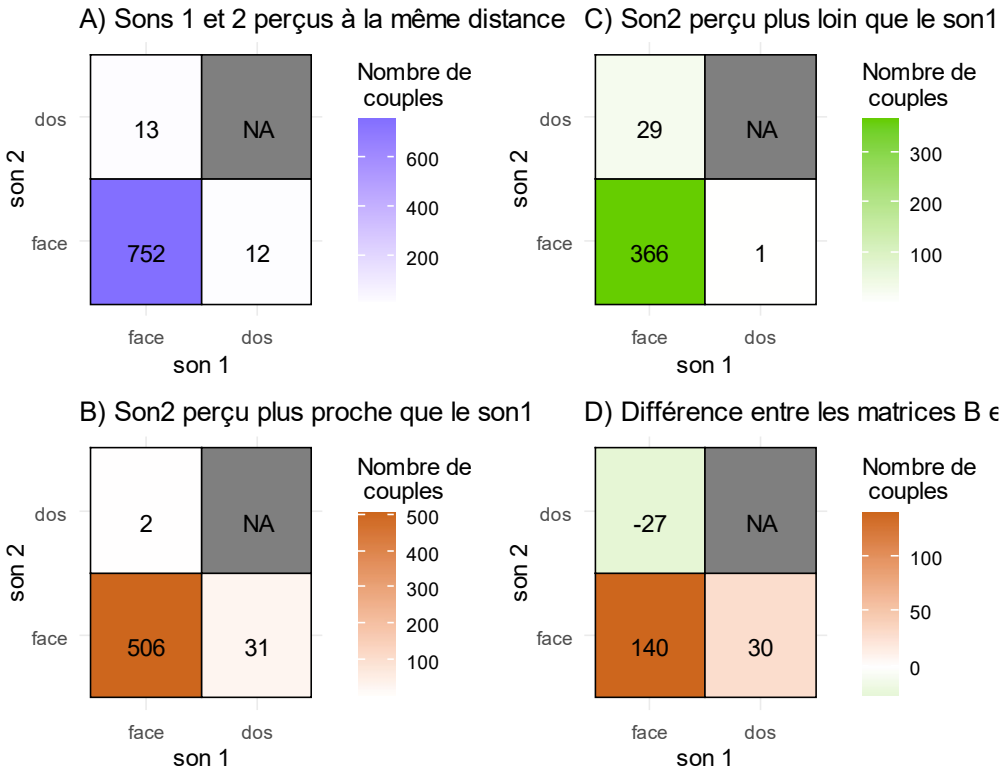


Figure 88 : Perception de la distance en fonction de l'orientation

Pour finir, rassemblons les résultats des 4 tests, et essayons de classer les étiquettes spatiales et vocales en fonction de la perception des sujets. Par exemple, regroupons tous les sons étiquetés « avant-proche », et calculons le nombre de fois où ils ont été perçus à la même distance que l'autre son du couple (=), plus proche (>), et plus loin (<). Étant données nos observations précédentes, on s'attend à ce groupe n'obtienne pas les mêmes pourcentages que le groupe « avant-loin ». En procédant de même pour tous les groupes d'étiquettes spatiales, on obtient le Tableau 36. On peut ensuite classer les éléments de ce tableau grâce à la méthode de Ward, qui consiste à regrouper les lignes les plus proches, tout en maximisant la distance entre chaque groupe. Le dendrogramme correspondant est représenté en Figure 91. Quatre groupes se démarquent : en particulier, on retrouve un groupe « dos », un groupe « proche / milieu_face » et un groupe « loin ». La seule exception concerne les sons venant de la gauche : ainsi, les sons « gauche-milieu-dos » ne sont pas classés avec le groupe « dos » (perçu à 70% plus loin), mais avec le groupe « loin » (perçus à 50% plus loin). Par ailleurs, les sons « gauche-milieu-face » et « gauche-loin » forment leur propre groupe, celui des sons perçus majoritairement à distance

fixe. On en déduit que les sujets ont eu plus de mal à deviner la distance des sons venant de la gauche. Ce résultat est surprenant, car a priori, les deux micros utilisés sont identiques, donc on devrait retrouver les mêmes résultats lorsque le son vient de la droite. Lorsqu'on étudie plus en détail les spectres moyens, on constate qu'effectivement les différences spectrales entre chaque distance sont moins bien marquées pour les sons venant de la gauche (cf. Figure 89) que pour les autres sons (cf. Figure 90).

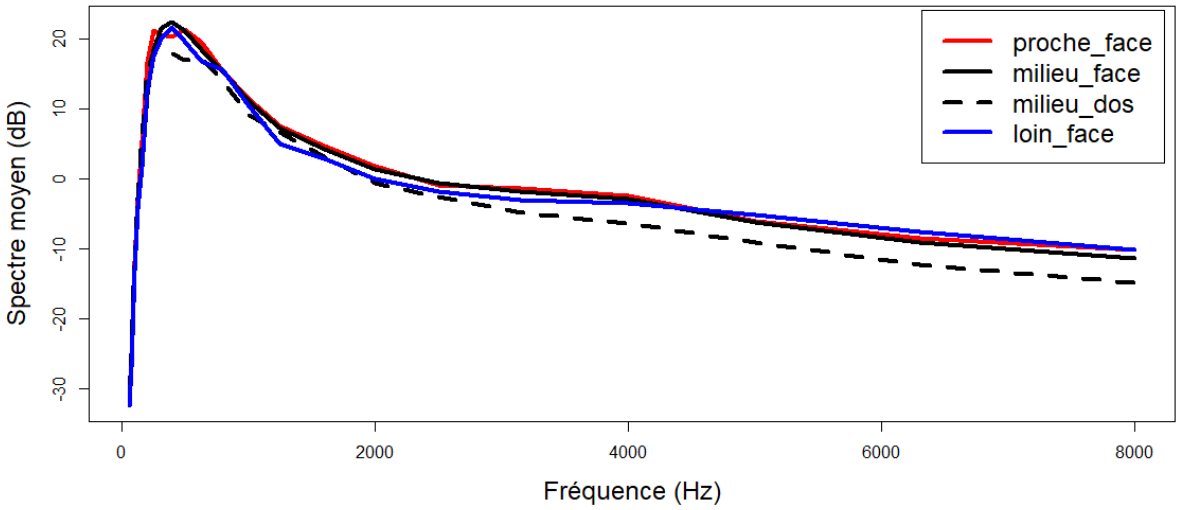


Figure 89 : Spectre moyen long terme pour les sons venant de la gauche (sons enregistrés par le robot)

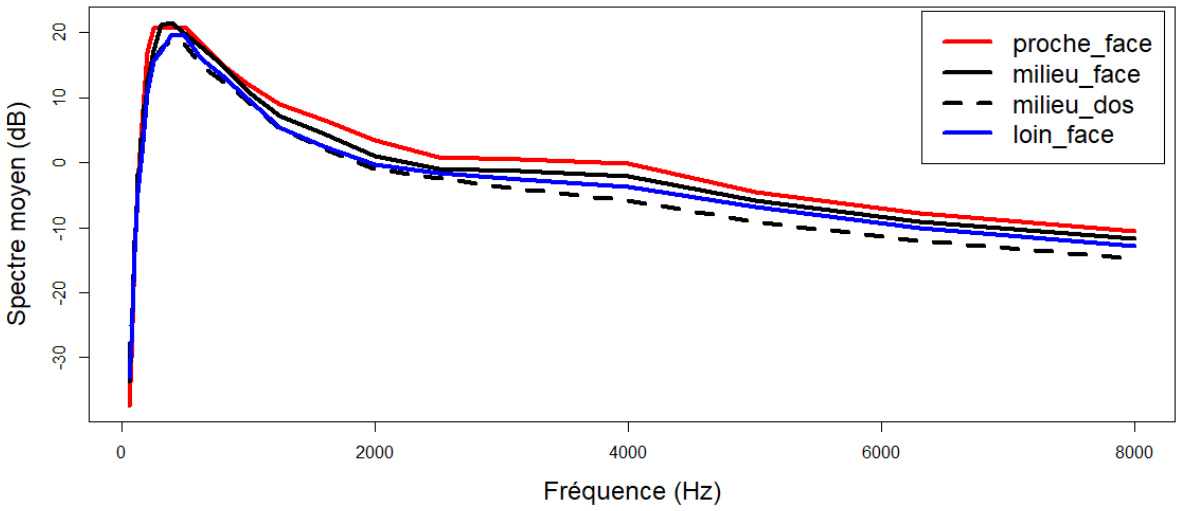


*Figure 90 : Spectre moyen long terme pour les sons **ne** venant **pas** de la gauche (sons enregistrés par le robot)*

Tableau 36 : Perception des sons en fonction de leurs étiquettes spatiales

direction	distance + orientation	=	>	<	nombre d'apparitions
avant	proche	43,1 %	44,4 %	12,5 %	304
avant	milieu_face	41,4 %	33 %	25,6 %	309
avant	milieu_dos	25 %	6,3 %	68,8 %	16
avant	loin	37,1 %	10,9 %	52,1 %	267
gauche	proche	51 %	34,2 %	14,7 %	292
gauche	milieu_face	59 %	21 %	20,1 %	329
gauche	milieu_dos	38,5 %	7,7 %	53,8 %	26
gauche	loin	54,6 %	15,1 %	30,3 %	317
droite	proche	46,6 %	42,4 %	11,1 %	262
droite	milieu_face	49,7 %	28,5 %	21,8 %	298
droite	milieu_dos	27,6 %	0 %	72,4 %	29
droite	loin	44,4 %	13,3 %	42,3 %	279
derriere	proche	37,7 %	46,2 %	16,2 %	247
derriere	milieu_face	39,5 %	30,4 %	30 %	263
derriere	milieu_dos	23,8 %	4,8 %	71,4 %	21
derriere	loin	43,3 %	15,9 %	40,8 %	245

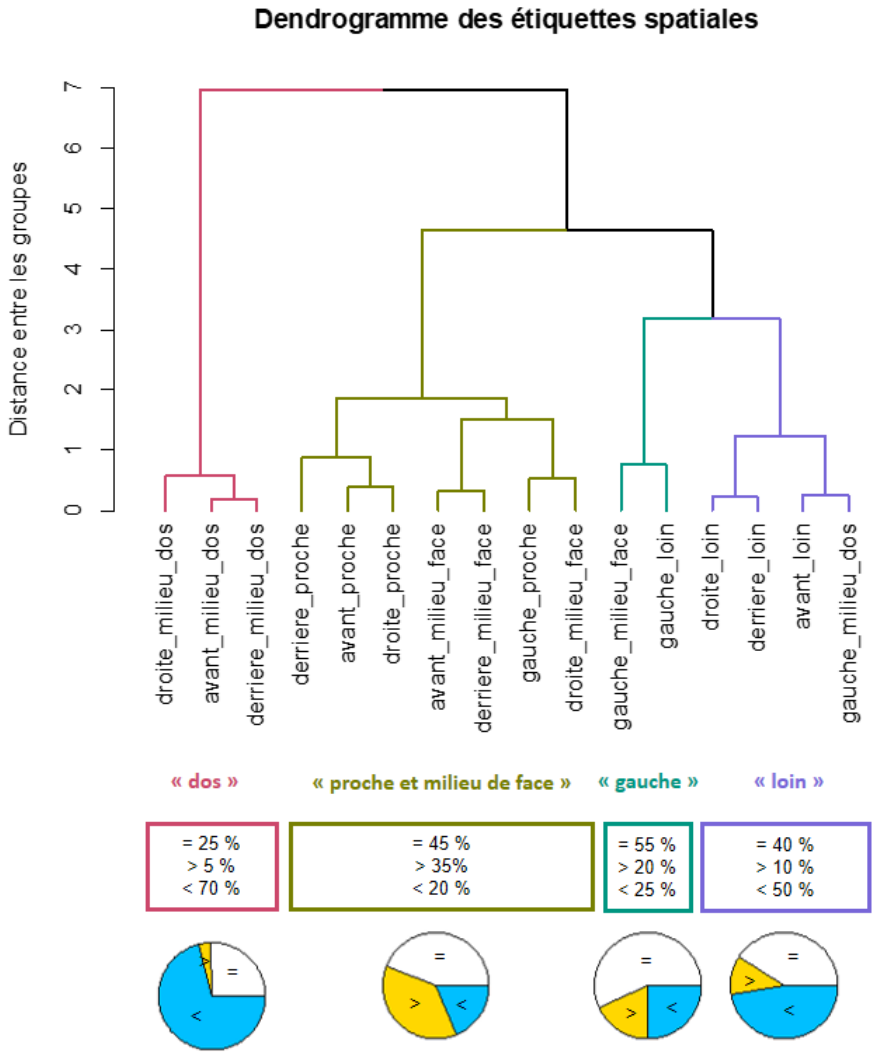


Figure 91 : Classification hiérarchique des étiquettes spatiales

On procède de la même manière, en classant cette fois les sons en fonction des étiquettes vocales (socio-affect, intensité et mot). Les résultats sont rapportés dans le Tableau 37, et le dendrogramme associé est visible en Figure 92. Cette fois, les groupes se démarquent en fonction du type de voix utilisée : « doute-faible », « doute-fort », « confiance-faible » et « confiance-fort ». À nouveau, on trouve des exceptions : ainsi, le groupe « doute-faible-coco » est classé avec les « confiance-faible », tandis que les groupes « doute-fort-ananas » et « doute-fort-brûlé » sont classés avec les « confiance-fort ». En outre, on observe clairement deux branches : une pour l'intensité faible, une pour l'intensité forte. C'est donc bien l'intensité qui domine. On remarque également que la part de sons classés dans la catégorie « neutre » (=) est plus importante lorsque l'intensité n'est pas cohérente avec le socio-affect.

En conclusion de ces tests sur la distance, l'influence de l'intensité apparaît prédominante. Ainsi, lorsque les sons du couple sont d'intensités différentes, c'est celui de plus forte intensité qui est perçu plus proche. De même, lorsque les sons du couple ne sont pas prononcés avec la même orientation, le son prononcé de face est perçu plus proche. Rappelons que dans les expériences précédentes, l'effet de l'intensité était beaucoup moins marqué : l'intensité influençait d'abord la perception de l'orientation, qui elle-même influençait la perception de la distance. Ici, non seulement les sujets étaient uniquement interrogés sur la distance, mais la tâche qui leur était demandée consistait à comparer deux sons : il s'agit donc d'une perception relative, et non d'une perception absolue comme précédemment. Dans ce cas, c'est donc principalement le contraste entre les deux sons qui a déterminé la perception des sujets. Le socio-affect a également une certaine importance, puisque pour une intensité donnée, la confiance est en moyenne perçue plus proche que le doute.

Tableau 37 : Perception des sons en fonction de leurs étiquettes vocales

socio-affect + intensité	mot	=	>	<	nombre d'apparitions
doute	brule	41,5 %	19,5 %	39 %	277
doute	coco	47,2 %	9,3 %	43,5 %	161
doute	champignon	44,3 %	21,2 %	34,4 %	273
doute	ananas	38,9 %	15,4 %	45,7 %	162
confiance	brule	47,7 %	17,9 %	34,4 %	279
confiance	coco	53 %	18,2 %	28,8 %	264
confiance	champignon	47,2 %	22,2 %	30,6 %	144
confiance	ananas	50,5 %	12,8 %	36,7 %	188
DOUTE	brule	40,5 %	39,5 %	20 %	220
DOUTE	coco	53,6 %	31,9 %	14,5 %	248
DOUTE	champignon	51,5 %	26,7 %	21,8 %	165
DOUTE	ananas	44,9 %	40,1 %	15 %	247
CONFIANCE	brule	43,1 %	39,9 %	16,9 %	248
CONFIANCE	coco	39,5 %	40 %	20,5 %	205
CONFIANCE	champignon	44 %	38 %	18 %	200
CONFIANCE	ananas	41,7 %	36,3 %	22 %	223

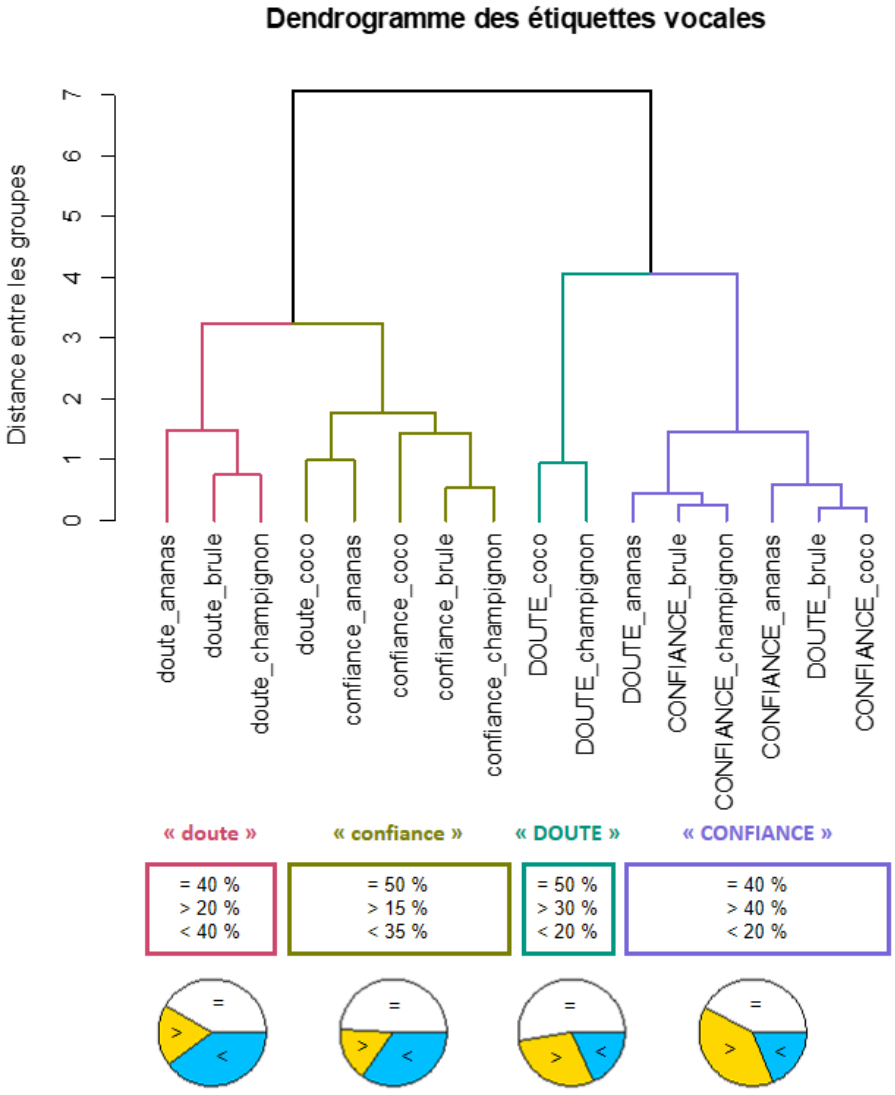


Figure 92 : Classification hiérarchique des étiquettes vocales

4.3.3.3 Perception du socio-affect (test 5)

Passons à présent au test 5, sur la perception du socio-affect. L'objectif de ce test était de vérifier que les sujets reconnaissaient bien le socio-affect utilisé, et éventuellement, de voir si le taux de reconnaissance pouvait être influencé par la position spatiale de la locutrice. Autrement dit, il s'agit de poser la question inverse de celle posée jusqu'à présent : Est-ce que la perception du socio-affect peut-être influencée par les variables spatiales ?

Afin de conserver les étiquettes utilisées aux expériences 1 et 2, les deux socio-affects étudiés sont encore désignés comme « doute » et « confiance », bien que les sujets les classent en réalité selon deux catégories : « poli / sympathique » et « autoritaire ». Les premiers résultats sont présentés au Tableau 38 sous la forme d'une matrice de confusion. Elle montre que les sujets arrivent très bien à classer les sons dans les deux catégories demandées : en moyenne, leur taux de reconnaissance s'élève à 87 %. À titre de comparaison, ce taux de reconnaissance moyen était de 78 % pour l'expérience 1, et de 89 % pour l'expérience 2. Cette fois, le doute est mieux reconnu que la confiance, contrairement à ce qu'on avait observé précédemment. Pour expliquer cette différence, on peut supposer qu'en cas d'hésitation, la plupart des sujets ont opté pour la réponse « voix polie / sympathique », plutôt que pour la réponse « voix autoritaire ».

Tableau 38 : Matrice de confusion pour le socio-affect

socio-affect		**Production**	
		doute	confiance
Perception	doute	485 **92 %**	112 21 %
	confiance	44 8 %	419 **79 %**

Dans la suite, on veut savoir si cette perception est influencée par les autres variables du test. Pour ce faire, nous allons étudier les résultats par sous-catégorie. Pour pouvoir calculer des écart-types, et ainsi juger de la significativité des résultats, nous calculerons le taux de reconnaissance obtenu par chaque sujet dans chaque sous-catégorie.

Le Tableau 39 présente les taux de reconnaissance obtenu en tenant compte de l'intensité du mot-clé. On retrouve des résultats observés dans l'expérience 2 : le taux de reconnaissance est particulièrement élevé dans le cas où le socio-affect est cohérent avec l'intensité. Ainsi, il monte à près de 99 % pour le doute faible et la confiance forte. La confiance faible en revanche est particulièrement mal reconnue en moyenne : seulement 64%, contre 86% pour le doute fort. Notons que les résultats varient fortement d'un sujet à l'autre, comme l'attestent les écarts-types présentés dans le tableau. Ainsi certains sujets parviennent à reconnaître 100% des socio-affects, tandis que d'autres obtiennent de très faibles résultats lorsque le socio-affect n'est pas

cohérent avec l'intensité. Les histogrammes détaillés sont présentés en Figure 93. Plus précisément, sur les 29 sujets, 8 obtiennent des taux de reconnaissance supérieurs à 80 % pour le doute fort et la confiance faible, 2 obtiennent un taux inférieur à 80 % uniquement pour le doute fort, 13 uniquement pour la confiance faible, et les 5 restants ont des taux de reconnaissance inférieurs à 80% pour les deux catégories.

Tableau 39 : Taux de reconnaissance du socio-affect en fonction du socio-affect et de l'intensité

Sous-catégorie	doute faible	doute fort	confiance faible	confiance forte
Taux de reconnaissance	**99,6 %** ± 2,4 %	**85,6 %** ± 18,6 %	**63,5 %** ± 28,0 %	**98,2 %** ± 5,6 %

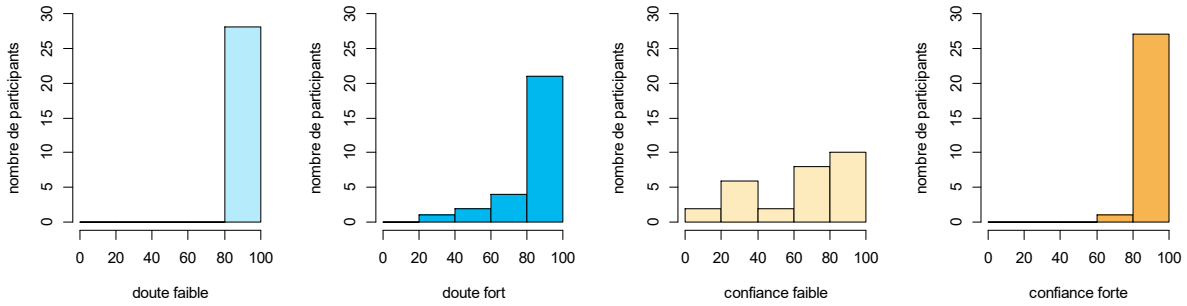


Figure 93 : Histogramme des taux de reconnaissance du socio-affect pour chaque sous-catégorie d'intensité (%)

Comme les taux de reconnaissance sont proches de 100% pour la confiance forte et le doute faible, il suffit donc de s'intéresser aux sous-catégories « doute fort » et « confiance faible » pour étudier l'influence des autres variables. Cependant, on arrive à une limite : malgré la trentaine de sujets ayant complété le test, le nombre de données reste trop faible, ce qui engendre des écarts-types très élevés sur le calcul des taux de reconnaissance. Le décompte séparé des erreurs est tout de même rapporté dans les Tableau 40, 35 et 36 (toutes les erreurs sont prises en compte, y compris celles des sujets qui n'ont pas fini le test). On constate en particulier que le nombre d'erreur pour la catégorie « confiance-faible » est légèrement plus faible lorsque le son est proche que pour les autres distances ; mais cette variation n'est pas significative (valeur-p : 0,3).

Tableau 40 : Décompte des erreurs en fonction de la distance et de l'orientation

Sous-catégorie	proche	milieu face	milieu dos	loin
doute faible	0	2	1	0
doute fort	11	16	10	12
confiance faible	19	30	31	30
confiance fort	0	5	2	3

Tableau 41 : Décompte des erreurs en fonction de la direction

Sous-catégorie	avant	gauche	droite	derrière
doute faible	0	1	0	2
doute fort	16	12	12	9
confiance faible	22	30	28	30
confiance fort	1	7	3	1

Tableau 42 : Décompte des erreurs en fonction du mot

Sous-catégorie	pin	rose	eucalyptus	fleur d'oranger
doute faible	2	1	0	0
doute fort	8	12	10	19
confiance faible	34	25	17	34
confiance fort	2	4	2	4

Plutôt que d'étudier uniquement les (rares) erreurs des sujets, on peut s'intéresser à leur temps de réponse. En effet, on peut supposer que lorsqu'une personne hésite entre deux réponses, elle met plus de temps à répondre. Au contraire, un temps de réponse court signifie qu'elle est certaine de son choix.

Malheureusement, l'histogramme des temps de réponse montre que les données sont loin de suivre une répartition normale (cf. Figure 94) : on observe un pic autour de 2 secondes, qui s'étale vers la droite ; on note également la présence de plusieurs valeurs extrêmes, qui biaiseraient fortement le calcul des moyennes. En effet, les sujets ne peuvent pas répondre plus vite que le temps nécessaire pour écouter le son. Pour analyser les données, nous avons donc choisi de supprimer les mesures supérieures à 10 000 ms (il n'y en a que 15 sur 896). Il est également possible de transformer les données pour les rapprocher de la distribution normale, par exemple en prenant le logarithme des mesures. Ici, nous avons choisi de conserver les données brutes afin de faciliter l'interprétation des résultats. En effet, nous avons vérifié que le modèle obtenu à partir des données transformées fournit les mêmes tendances que le modèle construit à partir des données brutes.

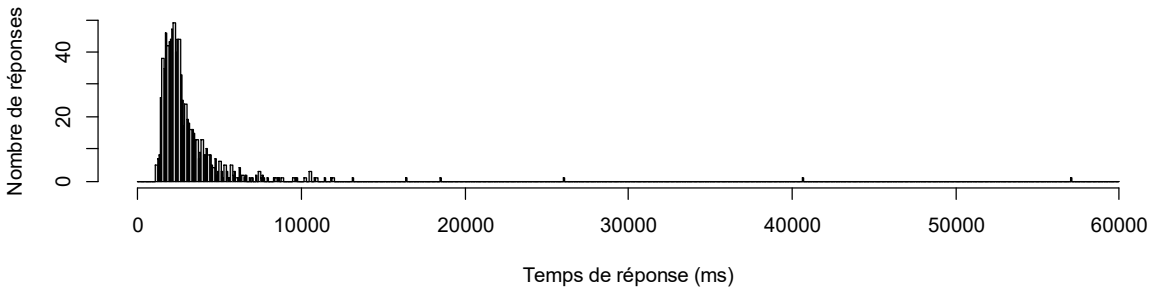


Figure 94 : Histogramme des temps de réponse

Cette fois, on ajoute la variable « mot » aux effets aléatoires. En effet, on constate que les sujets répondent un peu plus vite pour les mots courts, en particulier les mots « pin » (cf. Tableau 43).

On ajoute également la variable « numéro » : elle prend des valeurs de 1 à 32, et correspond au numéro du son dans le test. En effet, on constate que le temps de réponse des sujets décroît lentement au cours du test (cf. Figure 95).

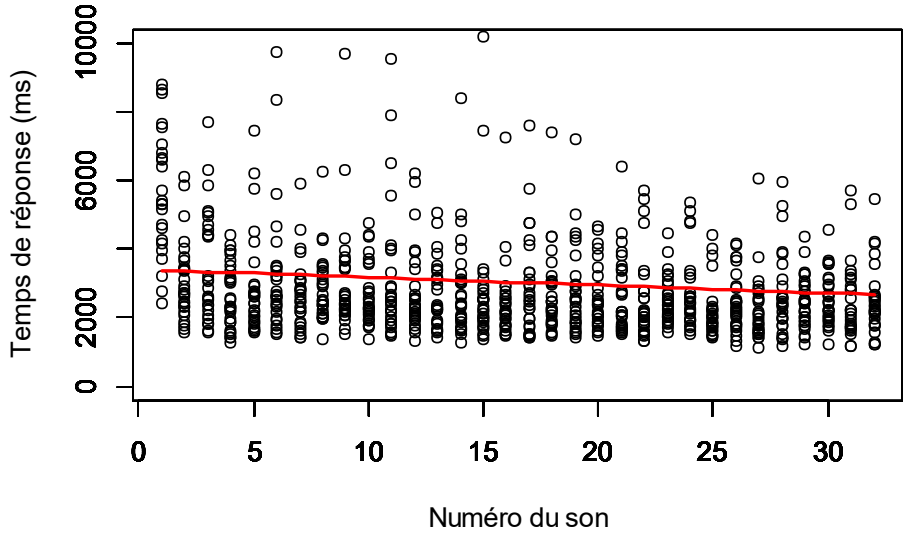


Figure 95 : Temps de réaction des sujets en fonction du numéro du son (modèle linéaire obtenu sans tenir compte de la première mesure)

Le modèle linéaire mixte s'écrit alors comme ceci :

$$\text{temps de réponse} \sim \text{type de voix} + (1 \mid \text{sujet}) + (1 \mid \text{mot}) + (1 \mid \text{numero})$$

Les résultats de la modélisation sont présentés dans le Tableau 44.

On constate que le temps de réponse des sujets est d'environ 15 % plus élevé pour les types de voix ambigus (doute fort et confiance faible), ce qui confirme notre intuition initiale. La p-valeur est de 2.10^{-10} .

Tableau 43 : Coefficients du modèle linéaire mixte modélisant le temps de réponse en fonction du mot entendu
$temps\ de\ réponse \sim mot + (1 \mid sujet) + (1 \mid type\ de\ voix) + (1 \mid numero)$

Direction	Temps de réponse
eucalyptus (intercept)	2893 ± 189 ms
fleur d'oranger	+ 78 ± 116 ms
pin	- 267 ± 115 ms
rose	- 116 ± 113 ms

Tableau 44 : Coefficients du modèle linéaire mixte modélisant le temps de réponse en fonction du type de voix utilisé
$temps\ de\ réponse \sim type\ de\ voix + (1\ |sujet)\ + (1\ |\ mot) + (1\ |\ numero)$

Direction	Temps de réponse
doute faible (intercept)	2713 ± 172 ms
doute fort	+ 487 ± 96 ms
confiance faible	+ 319 ± 96 ms
confiance fort	- 98 ± 95 ms

Pour étudier l'effet de la distance sur les temps de réponse, on tient donc compte de tous les effets aléatoires précédents :

$$\text{temps de réponse} \sim \text{distance} + (1\ |\text{sujet})\ + (1\ |\ \text{mot}) + (1\ |\ \text{numero}) + (1\ |\ \text{type de voix})$$

Les résultats de la modélisation sont présentés dans le Tableau 45. Ils ne sont pas significatifs (p-valeur : 0,1), bien qu'on constate que le temps de réaction modélisé est plus élevé pour les sons « loin » que pour les sons « proches ».

Tableau 45 : Coefficients du modèle linéaire mixte modélisant le temps de réponse en fonction de la distance
$temps\ de\ réponse \sim distance + (1\ |sujet)\ + (1\ |\ mot) + (1\ |\ numéro) + (1\ |\ type\ de\ voix)$

Direction	Temps de réponse
proche (intercept)	2733 ± 217 ms
milieu face	+ 66 ± 100 ms
milieu dos	+ 117 ± 100 ms
loin	+ 186 ± 100 ms

De la même manière, on peut montrer que la direction d'arrivée des sons n'a pas d'effet significatif. Cependant, on remarque qu'une part importante de la variabilité des données reste non expliquée par ces modèles, ce qui pourrait expliquer leur absence de significativité (cf. Tableau 46).

Tableau 46 : Contributions des effets aléatoires des modèles linéaires mixtes modélisant le temps de réponse

	Variance estimée	Erreur-type
sujet	386 103	552
numéro	302 388	621
type de voix	65 225	255
mot	9 936	102
Résidu	1 065 976	1034

Notons par ailleurs qu'à ce stade de l'expérience, les sujets sont déjà familiarisés avec les types de voix utilisés, puisqu'ils ont déjà eu l'occasion de les entendre au cours du test sur la distance. Il n'y a donc pas d'effet d'apprentissage : même si le temps de réponse diminue légèrement au cours du test, en moyenne, leurs résultats sont aussi bons au début et à la fin du test. Le test est également suffisamment court pour éviter que les sujets se fatiguent, et commencent à répondre au hasard. Pour le prouver, nous nous sommes intéressés à l'ordre dans lequel les erreurs se produisaient, afin de savoir si elles avaient lieu plutôt au début, en fin d'expérience, ou à n'importe quel moment. La Figure 96 représente ainsi les réponses d'un sujet sous forme de bulles, numérotées dans l'ordre d'écoute de 1 à 32 : sur la ligne du haut apparaissent les bonnes réponses, et sur la ligne du bas, les mauvaises réponses. Leur couleur correspond au type de voix utilisée. On constate sur cet exemple que le sujet ne s'est pas trompé de façon systématique : au moment où les erreurs se produisent, il a déjà reconnu correctement des « confiance faible » et des « doute fort », et en reconnaîtra encore avant la fin de l'expérience. Ce résultat se confirme pour l'ensemble des sujets : en moyenne, le numéro des sons mal reconnus est de 16,4 ce qui est très proche de la médiane ; il y a donc autant de sons mal reconnus au début du test qu'à la fin.

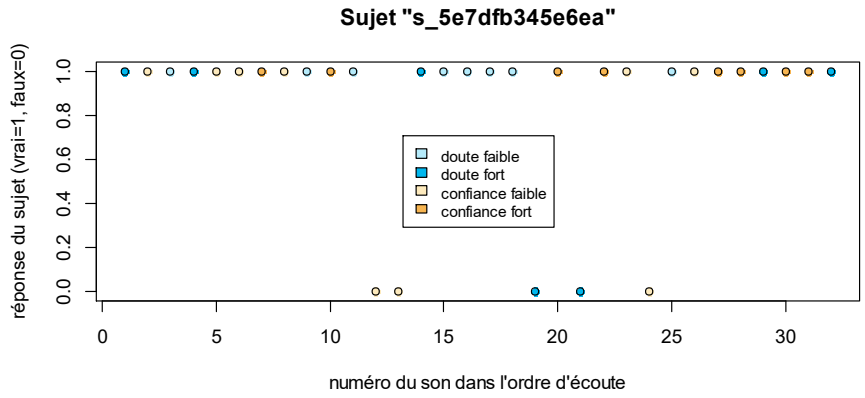


Figure 96 : Représentation graphique des réponses d'un des sujets

4.3.3.4 Perception de la direction (test 6)

Enfin, le test 6 consistait à écouter les sons un par un, et à reconnaître leur direction. 32 sujets y ont participé. Deux d'entre eux avaient visiblement inversé le sens de leur casque, nous avons donc corrigé leurs données en conséquence, en échangeant les étiquettes de direction d'arrivée des sons.

En moyenne, le taux de reconnaissance pour la direction est de 57%. Sans surprise, ce taux est beaucoup plus faible que ceux obtenus lors des expériences 1 et 2 ; mais il reste supérieur au hasard (25%). La matrice de confusion présentée dans le Tableau 47 montre que les sujets n'ont pas su distinguer si le son venait de devant ou de derrière le robot. Ce résultat est prévisible, car les auditeurs n'ont accès à aucun indice acoustique leur permettant de faire la distinction. Quatre d'entre eux ont d'ailleurs mentionné cette difficulté en commentaire. Comme observé précédemment, ils ont tendance à privilégier la direction avant. En regroupant les catégories

avant et derrière, le taux de reconnaissance monte à 78%. On remarque que pour les sons enregistrés en avant du robot, la part des sons perçus à gauche est plus importante que ceux perçus à droite, et inversement lorsque le son vient de derrière. Cette observation est cohérente avec les analyses du paragraphe 4.3.2 : on a constaté que l'écart d'intensité entre l'oreille gauche et l'oreille droite du robot est positif lorsque le son vient de l'avant, comme si le son venait de la gauche, et négatif lorsqu'il vient de derrière, comme si le son venait de la droite.

La gauche et la droite sont relativement bien reconnues, avec un taux de reconnaissance de plus de 70%. Il est également très rare que ces deux directions soient confondues (moins de 2% des cas).

Tableau 47 : Matrice de confusion pour la perception de la direction
(les nombres en bleu correspondent au nombre de données recueillies)

direction		**Production**			
		avant	gauche	droite	derrière
Perception	avant	265 **52 %**	55 11 %	61 12 %	275 53 %
	gauche	81 16 %	399 **77 %**	10 2 %	30 6 %
	droite	17 3 %	8 1 %	371 **72 %**	70 14 %
	derrière	151 29 %	55 11 %	73 14 %	141 **27 %**

La variabilité inter-sujets est élevée : certains obtiennent des taux de 100% pour la reconnaissance des sons venant de la gauche ou de la droite, tandis que d'autres ont beaucoup plus de difficulté. On peut supposer que leur matériel d'écoute ne leur permet pas de séparer correctement le son arrivant à l'oreille droite, et celui arrivant à l'oreille gauche. Les histogrammes détaillés sont fournis en Figure 97.

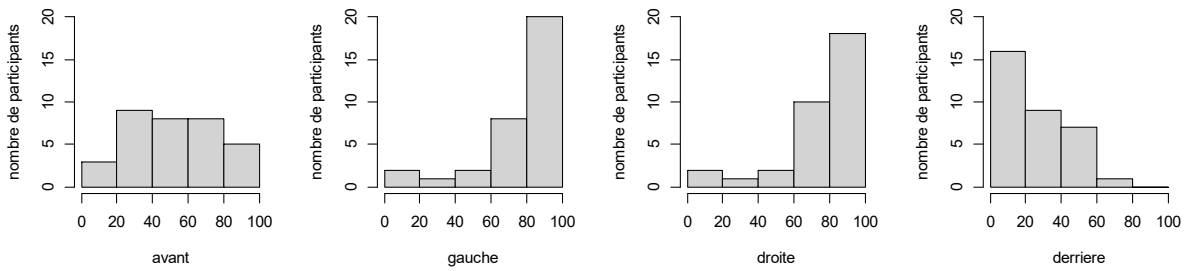


Figure 97 : Histogramme de taux de reconnaissance de la direction (%)

Regardons à présent plus en détails comment les participants décident si le son vient de l'avant ou de derrière. Nous avons représenté en Figure 98 les classes auxquelles appartiennent les sons perçus comme provenant de l'avant ou de derrière. En tout, 656 sons ont été perçus comme provenant de l'avant, et 420 comme provenant de derrière. On constate des écarts significatifs entre les répartitions, confirmés par un test du Khi-2 (cf. Tableau 48).

Ainsi, en moyenne, les sons perçus comme venant de derrière ont été plus souvent produits à une distance plus grande, de dos, ou avec une faible intensité que les sons perçus comme venant de l'avant. En revanche, il n'y a pas de différence significative pour le socio-affect, même si la tendance va à nouveau dans le sens de l'intensité : ainsi, 52 % des sons perçus de face sont des « confiance », et 52 % des sons perçus de dos sont des « doute ».

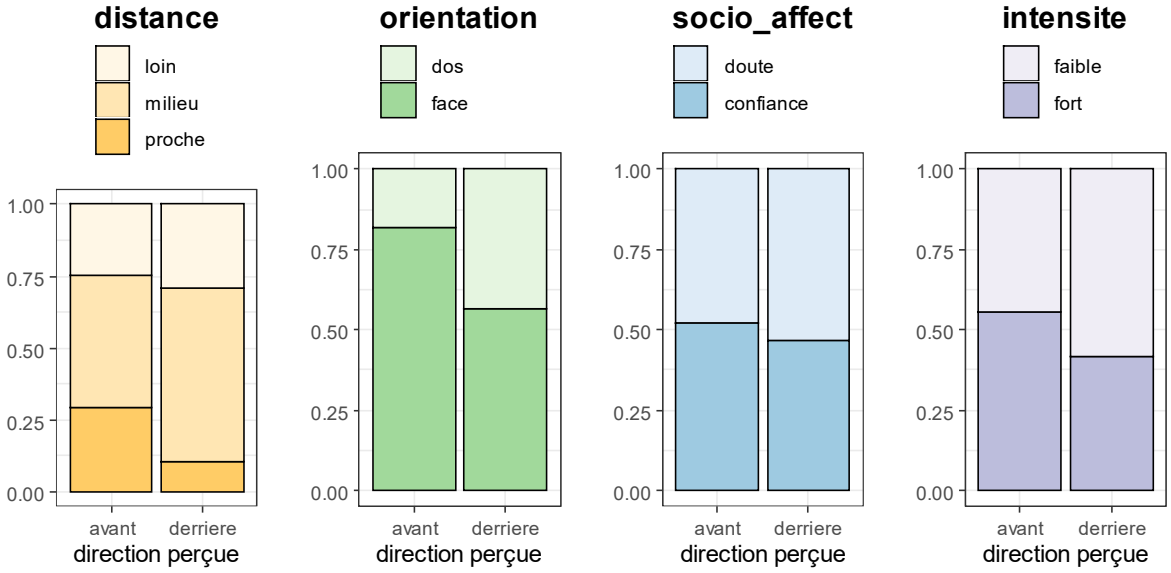


Figure 98 : Répartition des sons perçus comme venant de l'avant ou de derrière

Tableau 48 : Résultat au test du khi-2 comparant les répartitions des sons perçus comme venant de l'avant ou de derrière

	distance	**orientation**	**socio-affect**	**intensité**
valeur-p	10^{-12}	10^{-19}	10^{-1}	10^{-5}
significativité	***	***		***

4.3.4 Résumé

Pour cette troisième expérience sur la perception sociale de l'espace, nous avons remplacé le test psychoacoustique *in situ* par un test en téléprésence. Cette fois, notre corpus de stimuli est constitué de mots préenregistrés, et soigneusement sélectionnés pour correspondre à la prosodie désirée. Le test lui-même a été divisé en plusieurs parties, de manière à étudier séparément la perception de la distance, du socio-affect, et de la direction.

Les tests sur la perception de la distance confirment une évidence : les sujets se fient principalement à l'intensité pour classer les sons au sein de chaque couple. Ainsi, les sons faibles ont été perçus significativement plus proches que les sons forts. On retrouve la même

tendance entre les sons prononcés de dos et ceux prononcés de face, car le changement d'orientation induit un changement d'intensité. En allant dans les détails, on observe cependant un effet du socio-affect : les sons étiquetés « doute » ont été plus souvent perçus plus loin que les sons étiquetés « confiance ». Cette observation est cohérente avec la différence d'intensité entre ces deux socio-affects : le « doute » est intrinsèquement plus faible que la « confiance ».

Le test sur la perception du socio-affect confirme que celui-ci est bien reconnu par les sujets. En particulier, le taux de reconnaissance ainsi que le temps de réponse sont plus élevés lorsque l'intensité produite est cohérente avec le socio-affect.

Le test sur la perception de la direction montre que les sujets arrivent bien à reconnaître d'où vient le son, à condition d'exclure les confusions avant / arrière. Il confirme également un résultat observé lors des tests *in situ* : lorsque les sujets ne savent pas si le son vient de l'avant ou de derrière, ils répondent généralement que le son vient de l'avant. Il semble à nouveau que le choix des sujets se fassent essentiellement à partir de l'intensité perçue : en moyenne, les sons perçus comme venant de derrière ont été enregistrés plus loin, de dos, ou lorsque l'intensité produite était faible.

4.4 Conclusion

Nous avons conçu trois expériences dans le but d'étudier un effet très faible et difficile à mettre en évidence : l'effet de la distance sociale sur la perception de la posture d'un locuteur, et en particulier de sa distance physique. Le principe de ces expériences consistait à étudier comment les socio-affects d'une locutrice influencent la manière dont un auditeur estime sa position dans l'espace (direction, distance et orientation), et à quantifier cette influence.

À travers l'exemple de deux socio-affects aux caractéristiques opposées (« doute » et « confiance »), nous avons rapidement constaté qu'il n'était pas possible de dissocier totalement intensité et socio-affect : en effet, la locutrice avait plus de mal à respecter les consignes de production lorsque l'intensité demandée n'était pas cohérente avec le socio-affect. Sachant que l'intensité est un indice prépondérant pour la perception de la distance, nous avons choisi d'ajouter une consigne d'intensité (« faible » ou « forte »), c'est-à-dire une variation volontaire de la portée vocale, pour tenter d'isoler l'effet du socio-affect. Quatre types de voix ont donc été étudiés : deux types « naturels » (« doute faible » et « confiance forte ») et deux types « artificiels » (« doute fort » et « confiance faible »).

Cette corrélation entre intensité et socio-affect a eu des conséquences directes sur la perception des sujets. Tout d'abord, le socio-affect était bien moins reconnu lorsque les consignes vocales de socio-affect et d'intensité n'étaient pas cohérentes. De plus, l'effet du socio-affect allait généralement dans le même sens que celui de l'intensité intrinsèque : ainsi, les sujets ont plus souvent associé l'intensité faible et le socio-affect « doute » à une orientation de dos, à une distance éloignée, et au fait que le son venait de derrière. Cet effet était plus subtil à mettre en évidence que celui de l'intensité, mais nous sommes parvenus à l'observer à plusieurs reprises aux cours des expériences 1 et 3.

En revanche, il ne semble pas y avoir de lien entre la distance sociale exprimée et la perception de la distance physique : le « doute » (voix intime) n'a pas été perçu significativement plus proche que la « confiance » (voix distante). Inversement, la distance de production n'a pas influencé significativement la perception du socio-affect. La perception des sujets est donc robuste de ce point de vue.

La conclusion la plus importante de cette étude est d'ordre méthodologique : les résultats de notre deuxième expérience montrent que malgré les difficultés engendrées par l'utilisation d'une tâche prétexte, celle-ci était absolument nécessaire pour éviter que les sujets s'adaptent aux variations vocales de la locutrice. Dans d'autres cas, détourner l'attention des sujets peut permettre d'éviter de surestimer l'effet des facteurs étudiés. C'est ce que suggèrent les résultats de l'expérience présentée dans le chapitre suivant.

Chapitre 5 :
ALTÉRATION DE LA PORTÉE VOCALE PAR EFFET LOMBARD EN TÉLÉPRÉSENCE

Nous avons vu que pendant une interaction en présentiel, un locuteur est généralement capable de contrôler sa portée vocale, grâce à la perception directe de son environnement et de la manière dont sa voix s'y propage. En revanche, il existe plusieurs raisons pour lesquels le pilote d'un robot de téléprésence ne parvient pas à réaliser ce toucher vocal à distance. Tout d'abord, l'immersion fournit par ces robots est loin d'être parfaite : en particulier, la personne ne peut pas entendre comment sa voix sonne dans l'environnement distant, ni percevoir correctement la distance qui la sépare de ses interlocuteurs. Même en supposant que les productions vocales de la personne téléprésente correspondent à ce qu'elle désire transmettre, rien ne garantit a priori que la voix qui sort du robot reproduise parfaitement celle de son pilote : en particulier, son intensité peut être différente. Enfin, on peut imaginer qu'en cas de téléprésence ubiquïte, la personne puisse être influencée par son propre environnement local : par exemple, si l'environnement local est bruyant, elle aura tendance à parler plus fort, même lorsque ce bruit n'est pas audible par ses interlocuteurs. Cette absence de proprioception à distance pourrait déranger les interlocuteurs, comme suggéré au paragraphe $381.2.2.4. De façon plus subtile, une variation de portée vocale inadaptée au contexte pourrait être interprétée comme une variation socio-affective ; en effet nous avons vu aux chapitres 2 et 4 qu'il est très difficile de séparer ces deux dimensions. Ainsi, imaginons le cas d'une personne en chambre d'hôpital, qui parle inhabituellement doucement dans son environnement local pour ne pas déranger ses voisins : à distance, elle pourrait être perçue comme triste ou fatiguée par ses interlocuteurs. Au contraire, une personne qui parlerait trop fort parce qu'elle entend mal ses interlocuteurs pourrait être perçue comme énervée.

Nous avons vu en section $1.4 que l'immersion acoustique est déjà un objet de recherche bien étudié, y compris en robotique de téléprésence. Par ailleurs, il suffit de quelques précautions pour s'assurer que la voix qui sorte du robot soit à peu près à la même intensité que celle du pilote : on peut utiliser un micro-casque rigide, pour que l'enregistrement se fasse à une distance fixe de la bouche du pilote, et calibrer le volume sonore du robot. Dans ce chapitre, nous avons voulu explorer spécifiquement la problématique posée par la situation d'ubiquïté à travers l'étude de l'effet Lombard en téléprésence. L'effet Lombard désigne les variations vocales observées en présence de bruit. En cas d'interaction *in situ*, tous les interlocuteurs produisent une parole Lombard, ce qui renforce leur intelligibilité. En revanche, en cas d'interaction à distance, il existe deux cas de figure : soit le bruit provient de l'environnement du robot, et est donc audible à la fois par le pilote et par les interlocuteurs ; soit le bruit provient de l'environnement distant, et dans ce cas, seul le pilote peut l'entendre. Dans ce cas, l'effet Lombard est parasite : c'est un effort inutile, qui pourrait même nuire à l'intelligibilité, s'il n'est pas correctement identifié par les interlocuteurs.

Le but de cette étude est donc d'évaluer si l'effet Lombard peut altérer le toucher vocal en téléprésence, et si oui, dans quelles proportions. Pour ce faire, nous avons proposé une nouvelle expérience, afin de comparer la voix du pilote du robot de téléprésence dans plusieurs conditions de bruit.

5.1 Effet Lombard

Lorsque l'environnement est bruyant, nous avons tendance à parler plus fort. Ce phénomène a été nommé effet Lombard, d'après les travaux du médecin Etienne Lombard (Lombard 1911). Cet effet n'est pas propre à l'être humain, et a été mis en évidence chez de nombreuses autres espèces animales, en particulier des oiseaux et des mammifères. Il s'agit d'une adaptation évolutive, qui permet à ces espèces de maintenir leur communication acoustique malgré le bruit (Zollinger, Brumm 2011).

Dans cette section, nous décrirons les modifications vocales qui caractérisent l'effet Lombard. Nous verrons en quoi ce phénomène peut être qualifié de semi-automatique, puis, nous nous intéresserons à la manière dont il a été étudié.

5.1.1 Modifications vocales en présence de bruit

Si l'augmentation de l'intensité vocale en présence de bruit est la manifestation la plus évidente de l'effet Lombard, il existe d'autres variations significatives entre parole standard et parole Lombard. En particulier, l'augmentation de l'intensité s'accompagne d'une augmentation de la fréquence fondamentale, mise en évidence notamment par (Boril, Pollák 2005).

Des modifications de timbre sont également observables : en particulier, un déplacement de l'énergie des basses fréquences (< 500 Hz) vers les moyennes et hautes fréquences (> 1.2 kHz), une variation de l'inclinaison spectrale (*spectral tilt*), ainsi qu'un décalage des formants, en particulier F1 et F2 (Junqua 1996). Ces observations rejoignent celles obtenues dans le cadre plus général des études sur l'effort vocal : ainsi, (Liénard 2018) obtient des résultats similaires, mais à partir d'un corpus obtenu en demandant simplement à des sujets de parler à différents niveaux d'intensité.

La prosodie est également impactée, avec en particulier un allongement de la durée des voyelles (ex. : Summers et al. 1988; Tartter, Gomes, et Litwin 1993; Newman 2003). À nouveau, un parallèle est possible avec les résultats de (Fux 2012), qui concernent un effort vocal produit non pas par effet Lombard, mais dans le but de communiquer à grande distance.

Par ailleurs, les recherches de (Garnier et al. 2018) suggèrent qu'en environnement bruyant, les locuteurs ont recourt à l'hyper-articulation : ils exagèrent leurs mouvements de lèvres, afin de renforcer les indices visuels disponibles pour leur interlocuteur.

Notons que ces variations ont un effet si significatif que la parole Lombard est mal reconnue par les systèmes de reconnaissance automatique, y compris les plus modernes (Marxer et al. 2018).

5.1.2 Nature de l'effet Lombard

La nature de l'effet Lombard a beaucoup été débattue : s'agit-il d'un simple réflexe physiologique, ou bien les locuteurs adoptent-ils des stratégies plus sophistiquées pour renforcer leur intelligibilité en présence de bruit ? La réponse à cette question est importante pour la robotique de téléprésence : dans le premier cas, l'effet Lombard serait inévitable ; dans le second cas, on pourrait imaginer que des personnes entraînées seraient capables de s'adapter à l'origine du bruit, pour n'en tenir compte que lorsque l'interlocuteur l'entend également.

Initialement, l'effet Lombard était considéré comme un phénomène involontaire : en effet, en présence de bruit, un locuteur peut modifier sa manière de parler sans réfléchir, voire même sans s'en rendre compte. L'effet Lombard serait donc un simple réflexe, permettant de compenser le retour autophonique des locuteurs dégradé en présence de bruit. C'est pourquoi, certains auteurs parlent de « réflexe de Lombard » (Zollinger, Brumm 2011). Un autre argument qui va dans ce sens est que l'effet Lombard est très difficile à inhiber consciemment. Ainsi, (Pick et al. 1989) obtiennent des succès très mitigés lorsqu'ils demandent à des locuteurs de contrôler leur intensité vocale : même aidés d'une interface visuelle, ils n'y parviennent pas entièrement.

Pourtant, toutes les études sur l'effet Lombard notent également une importante variabilité inter-sujets : si tous augmentent leur intensité vocale en présence de bruit, l'intensité maximale, elle, varie d'un sujet à l'autre. En effet, augmenter sa force de voix est une stratégie très coûteuse énergétiquement pour les locuteurs[26], qui disposent d'autres manières pour renforcer leur intelligibilité. Ces stratégies alternatives ne constituent pas un simple réflexe, mais supposent l'implication de processus cognitifs de plus haut niveau. (Garnier 2007) identifie ainsi sept stratégies principales permettant de renforcer l'intelligibilité :

- augmentation de l'audibilité (en parlant plus fort)
- augmentation des contrastes acoustiques entre le bruit ambiant et la parole (au niveau temporel et spectral)
- facilitation de la recherche d'informations (en parlant plus lentement et en mettant l'accent sur certains mots)
- augmentation des indices phonétiques (contraste segmental)
- réduction de la charge mentale (mots plus simples, phrases plus courtes)
- redondance (répétitions)
- multi-modalité (utilisation d'indices visuels)

Une partie des modifications acoustiques observées dans la parole Lombard proviendrait donc d'un apprentissage et/ou d'une prise de décision du locuteur, qui s'adapte au contexte en

[26] Selon (Lindblom 1990), la variabilité de la parole peut s'expliquer par un simple principe du moindre effort : il existerait un continuum, allant de l'hypo-articulation à l'hyper-articulation.

fonction de sa personnalité, de son état physique et mental, ou encore de ce qu'il perçoit de ses auditeurs (cf. Figure 99).

Ces deux approches ne sont pas mutuellement exclusives : ainsi, (Luo et al. 2018) proposent un modèle neurologique dans lequel l'effet Lombard est un réflexe, mais influencé par des processus cognitifs de haut niveau.

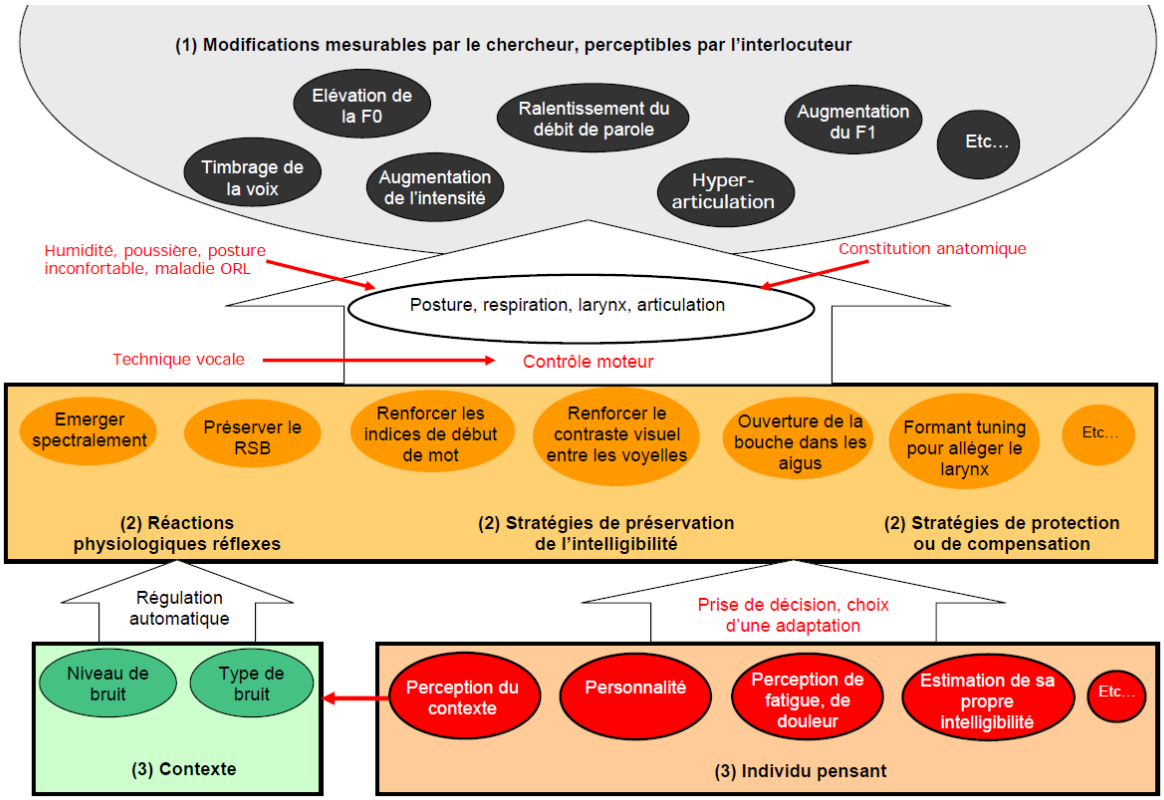


Figure 99 : Modélisation de l'adaptation de la parole dans le bruit (extraite de (Garnier 2007), p 215)

5.1.3 Limites des études antérieures

L'effet Lombard a été étudié dans de multiples conditions : avec différentes types de bruit et différents niveaux d'intensité. Cependant, ces conditions ne sont pas très adaptées pour étudier l'effet Lombard dans le cas particulier de la robotique de téléprésence, comme nous allons le voir en étudiant les graphiques présentés en Figure 100. Ces graphiques ont été obtenus à partir de la synthèse bibliographique proposée dans l'Annexe Bib2 de la thèse de (Garnier 2007). En voici une analyse détaillée :

a) On constate que dans la majeure partie des études, la tâche demandée aux sujets est très artificielle : dans 70% des cas, il s'agit de lire un texte à voix haute. Seules 15% des études ont tenté de générer de la parole spontanée, parfois en demandant simplement aux sujets de monologuer sur un sujet de leur choix.

b) Sur les 41 études, publiées entre 1954 et 2006, seules 14 mentionnent la présence d'un auditeur au moment de l'enregistrement. Cela signifie que dans au moins 62% des cas, le

locuteur parle seul, dans une pièce vide. Comme il n'y a pas d'interaction, les sujets n'ont pas besoin de se faire entendre. Ils n'ont pas non plus de retour visuel ou acoustique, qui leur permettrait de savoir si leur voix est audible par un interlocuteur.

c) Dans les trois-quarts des études, le bruit entendu par les sujets est diffusé au casque, ce qui permet de séparer facilement la parole des sujets du bruit, et facilite ainsi les mesures. Cependant, le port du casque modifie la manière dont le locuteur entend sa propre voix : un feedback audio est donc parfois utilisé pour compenser la diminution d'intensité liée au port du casque. Dans 15% des études, les sources sonores sont des haut-parleurs. Seules 10% des études se déroulent *in situ*, ou reproduisent en laboratoire des conditions *in situ,* c'est-à-dire en reproduisant les conditions acoustiques d'un lieu particulièrement bruyant, tel que les lieux de restauration ou les salles de classe.

d) Les bruits utilisés sont majoritairement artificiels : il s'agit de bruit blanc, rose, ou filtré pour avoir la même enveloppe spectrale qu'un bruit de conversation, ou n'occuper que les hautes ou les basses fréquences. Dans 33% des cas, il s'agit de bruits enregistrés, par exemple des conversations. Parfois, le bruit a été enregistré dans un lieu particulier : voiture, salle de classe, bar, etc. Seules 8% des études utilisent des bruits « naturels » : bruit de conversation entre des locuteurs en chair et en os, ou bruit de trafic routier par exemple.

e) Enfin, les bruits utilisés sont souvent très intenses, jusqu'à des niveaux dangereux en cas d'écoute prolongée (HCSP 2013). Dans des conditions si extrêmes, des personnes souhaitant communiquer préfèreraient d'abord s'éloigner des sources de bruit, ou attendre que le bruit passe : l'effort vocal qui leur est demandé n'est donc pas réaliste, à moins de considérer des cas très particuliers, tel que celui des pilotes d'avion.

5.1.4 Résumé

L'effet Lombard est donc un phénomène provoqué par la présence de bruit. Il s'agit principalement d'une augmentation de la force de voix, qui se traduit acoustiquement par une élévation de l'intensité, de la fréquence fondamentale, et du timbre de la voix. Ce réflexe, qui permet de renforcer l'intelligibilité de la voix, s'accompagne d'adaptations plus complexes, qui apparaissent au niveau prosodique, voire même linguistique. Cependant, il a principalement été étudié dans des conditions de laboratoire très éloignées de celles auxquelles on peut raisonnablement s'attendre en robotique de téléprésence. Pour pouvoir prévoir l'impact de l'effet Lombard sur la portée vocale en téléprésence, nous avons donc conçu une nouvelle expérience, en suivant les principes méthodologiques annoncés au chapitre 3. En particulier, nous avons cherché à mesurer un effet Lombard sans que les sujets aient conscience que le bruit faisait partie de l'expérience, afin que leur comportement soit similaire à celui qu'ils auraient face à un bruit ambiant parasite.

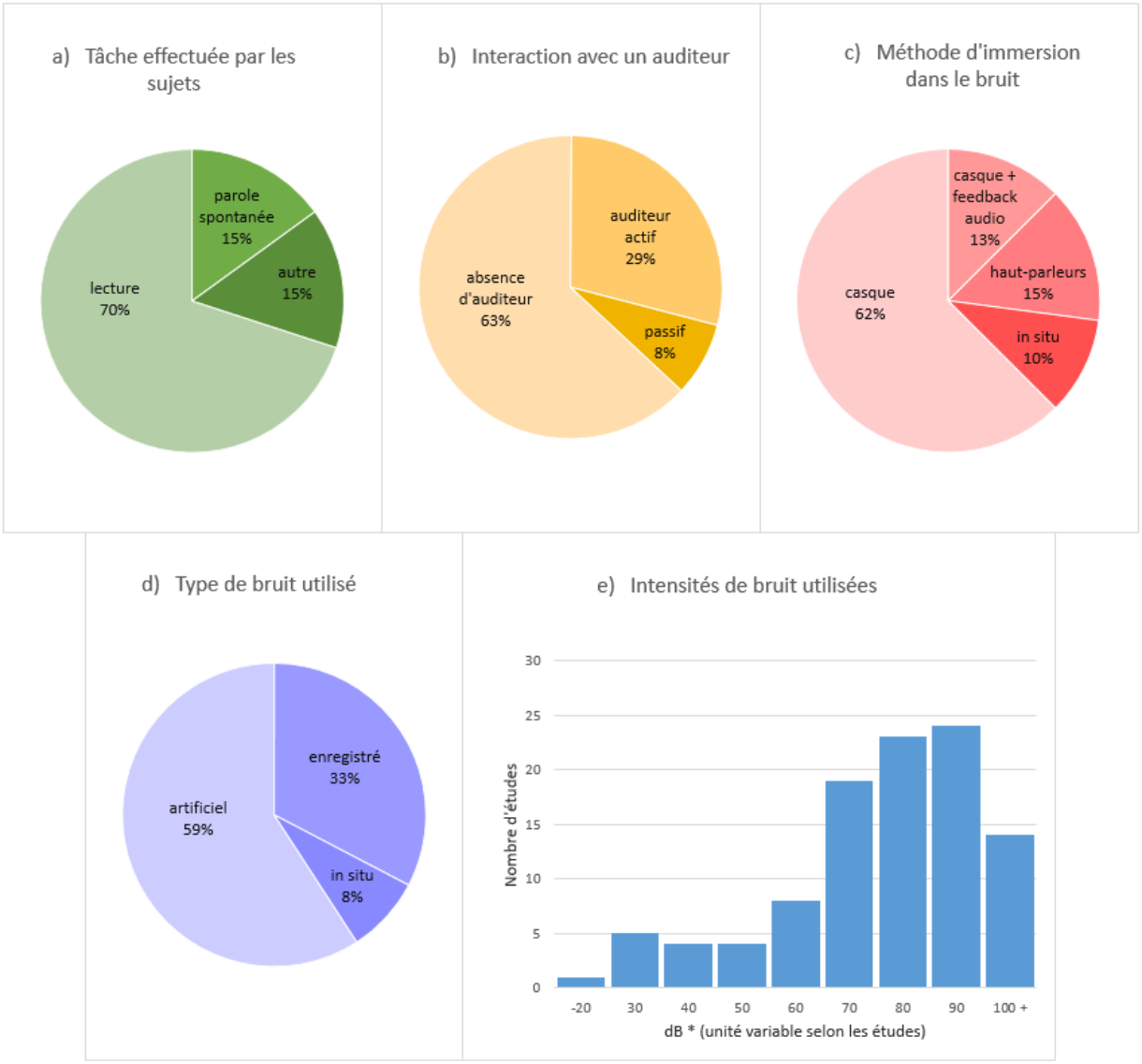


Figure 100 : Fréquences des études sur l'effet Lombard en fonction :
a) de la tâche donnée aux sujets b) de la méthode d'immersion c) du type d'interaction étudié
d) du type de bruit utilisé c) des intensités de bruit utilisées.

D'après la synthèse bibliographique en Annexe 2 de la thèse de (Garnier 2007), p 240-247

5.2 Nouvelle expérience

Dans cette partie, nous présenterons en détails le protocole expérimental que nous avons choisi pour étudier l'effet Lombard dans le cas d'une interaction en téléprésence. Nous avons fait en sorte que l'intensité des signaux acoustiques soit conservée entre l'environnement distant et l'environnement local. De plus, afin de limiter les variations d'intensité liée à la distance entre le microphone et la bouche du pilote, celui-ci est équipé d'un micro-casque.

5.2.1 Objectif et méthodologie

Lorsque le pilote d'un robot de téléprésence entend un bruit, ce bruit peut provenir soit de son propre environnement, soit de l'environnement distant (entendu au casque). Dans un cas, seul le pilote est capable d'entendre le bruit. Dans l'autre, le pilote et son interlocuteur entendent tous deux le même bruit. Nous avons souhaité comparer la parole produite dans ces deux conditions, à celle produite en l'absence de bruit. Nous avions également l'intuition qu'il serait possible d'observer un effet d'entraînement : c'est-à-dire qu'au cours d'une interaction en présence de bruit, le sujet s'adapterait à la fois au bruit ambiant, mais également aux variations d'intensité de son interlocuteur. Quatre situations étaient donc prévues initialement :

a) Situation de contrôle : pas de bruit
b) Présence de bruit dans l'environnement du pilote
c) Présence de bruit dans l'environnement du robot sans effet Lombard chez l'expérimentateur
d) Présence de bruit dans l'environnement du robot avec effet Lombard chez l'expérimentateur

Cette expérience aurait eu pour objectif secondaire de tester l'immersion acoustique à distance pour le pilote du robot de téléprésence. Ainsi, la consigne donnée au sujet aurait été de repérer dans l'espace l'expérimentatrice. Quelques essais sur le plateau étaient prévus pour expliciter les consignes et mesurer l'intensité produite par le sujet lorsqu'il se trouve dans la même pièce que l'expérimentatrice. Ensuite, le sujet aurait été conduit à la salle de pilotage, pour continuer le test en téléprésence. Les différentes situations de bruit auraient été mises en place de façon aléatoire au cours de l'expérience.

Ce protocole initial a évolué plusieurs fois avant d'arriver à sa version finale. Ces évolutions sont décrites dans le rapport de (Le Goff, Giorgis 2019). Finalement, nous avons choisi de faire appel à deux sujets afin d'observer des interactions plus spontanées et réalistes. En effet, il aurait été dommage de ramener l'étude de l'effet Lombard à une situation très artificielle dans laquelle un des deux sujets a conscience des buts de l'expérience, et sait ce que l'autre sujet entend ; d'autant plus que la téléprésence est un cas concret dans lequel deux interlocuteurs peuvent avoir une perception différente d'un même bruit.

Les quatre situations testées ont donc évolué pour devenir celles-ci :

A) Situation de contrôle : pas de bruit
B) Présence de bruit dans l'environnement du pilote
C) Présence d'un bruit « virtuel » provenant de l'environnement du robot
D) Présence de bruit dans l'environnement du robot

Ces quatre conditions correspondent bien aux quatre précédentes, à la différence près que cette fois l'interlocuteur est également un sujet de l'expérience. Ainsi, dans la condition D, l'interlocuteur local entend le bruit et doit donc être affecté par l'effet Lombard. De même, la condition C correspond à un bruit diffusé directement dans le casque du pilote : c'est comme si le bruit venait de l'environnement du robot, mais que l'interlocuteur n'était pas affecté par l'effet Lombard. Dans la suite, les deux sujets seront désignés par les initiales P et R : P pour le pilote, et R pour l'interlocuteur face au robot.

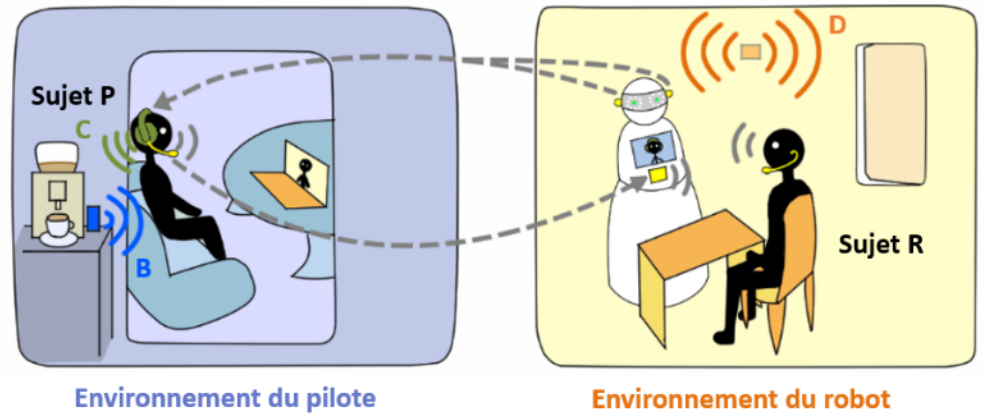


Figure 101 : Schéma de l'expérience sur l'effet Lombard en téléprésence

5.2.2 Prétexte

Les sujets étaient invités à participer à une expérience permettant d'évaluer l'ergonomie de l'interface de notre robot de téléprésence. Il était précisé que l'expérience se déroulait en binôme et que les sujets ne devaient pas avoir de problème d'audition ou de vue non corrigé. A leur arrivée, ils étaient conduits sur le plateau expérimental pour une courte introduction : rappel de ce qu'est un robot de téléprésence et explication de la tâche. Un des sujets était désigné pour poser une liste de questions au pilote. Toutes les dix questions, une consigne invitait le pilote à manipuler l'interface du robot. L'objectif annoncé était de comparer le temps mis pour répondre aux questions, et le temps nécessaire pour répondre aux consignes. Certaines consignes se répétant au cours de l'expérience, il serait possible d'évaluer la prise en main de l'interface. Pour éviter que les sujets interrompent leur échange pendant les bruits, ils avaient pour consigne de parler de façon régulière, sans s'arrêter, soit disant pour simplifier la mesure de leurs temps de réponse.

5.2.3 Choix des bruits

Nous avons choisi d'utiliser des bruits du quotidien, afin que les sujets pensent que ces bruits étaient accidentels et ne faisaient en aucun cas partie de l'expérience. En outre, nous voulions que les sujets P soient capables d'identifier si le bruit provenait de leur environnement, ou de l'environnement distant. C'est pourquoi nous avons utilisé deux sources de bruit différentes : une produisant des bruits de machine à café pour la condition B, et une produisant des bruits de perceuses pour la condition D. En pratique, certains sujets P n'ont pas reconnu le bruit de la machine à café, et n'ont donc pas toujours compris d'où venait ce son.

Tous les bruits ont été enregistrés à Domus à l'aide d'un enregistreur Zoom H6. Les bruits de machine à café proviennent de la véritable machine à café représentée en Figure 1. Les bruits de perceuse ont été enregistrés à l'occasion de travaux effectués dans le bâtiment. Afin que les bruits ne se répètent pas à l'identique, ce qui aurait pu éveiller la suspicion des sujets, il y avait 5 extraits de machine à café et 3 bruits de perceuse différents. Chaque extrait était très court : entre 5 et 11 secondes.

Par ailleurs, le volume des deux sources de bruits a été réglé de façon à mesurer environ 55 dB(A) au niveau des sujets P et R. Il s'agit du volume maximum que nous pouvions obtenir pour le bruit D en utilisant les haut-parleurs Bluetooth à notre disposition. C'est un niveau de bruit relativement faible comparé à ceux utilisés dans les études sur l'effet Lombard. En effet, dans ces études, le niveau de bruit descend rarement en dessous de 60 dB(A) et peut monter jusqu'à 100 dB(A) (cf. Figure 100). Il s'agit toutefois d'un niveau de bruit réaliste pour une application en robotique de téléprésence. En effet, il est rare de commencer une visioconférence ou un appel téléphonique dans un environnement bruyant quand on peut l'éviter.

Deux exemples de bruit utilisés sont présentés en Figure 102. Il s'agit non pas des enregistrements bruts diffusés pendant l'expérience, mais d'enregistrements effectués avec le robot de téléprésence. Pour le bruit B, le robot était placé dans la chambre sourde, à la place occupée habituellement par les sujets P. Pour le bruit C et D, le robot était placé sur la plateforme expérimentale, comme pendant l'expérience. Ces enregistrements sont stéréo, mais par soucis de clarté, seul le canal droit (le plus proche de la source de bruit) est représenté ci-dessous.

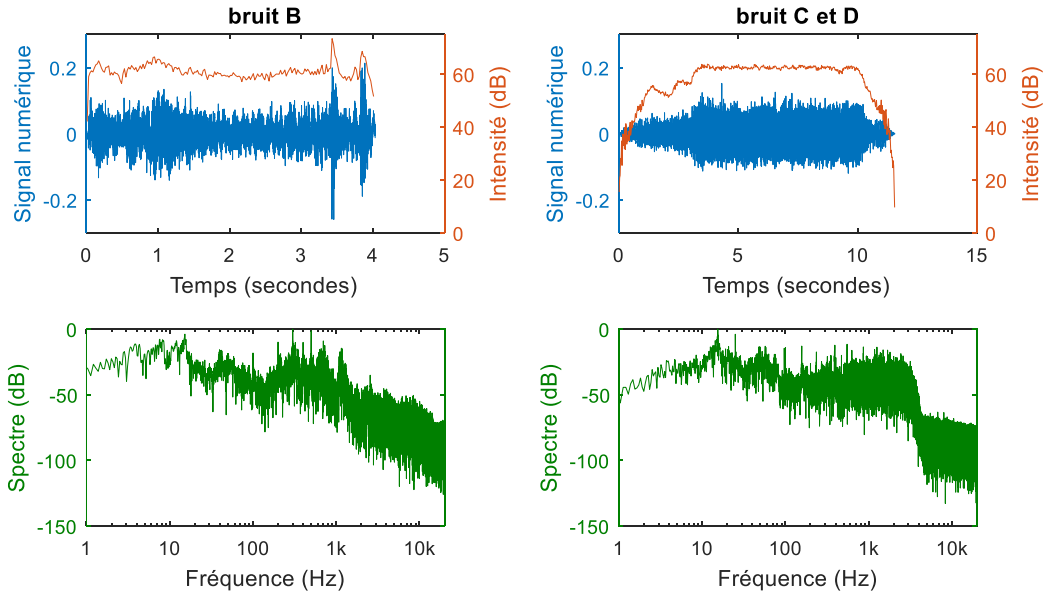


Figure 102 : Exemples de bruits utilisés

5.2.4 Choix de la tâche

La principale difficulté pour ce type d'expérience est d'avoir des mesures comparables entre elles. En effet, la mesure de l'intensité dépend fortement du contenu de l'extrait choisi : plus l'extrait est court, plus la mesure varie. Ainsi, deux mots d'une même phrase n'ont pas la même intensité moyenne, du fait de la prosodie de la phrase et du contenu phonétique de chaque mot. Il a donc fallu choisir une tâche permettant d'obtenir des productions comparables entre elles.

Nous avons opté pour une tâche de questions / réponses, afin que les productions du sujet P soient les spontanées possibles. La liste des questions a été rédigée par Emeline Le Goff et Zoé Giorgis (cf. Annexe D). Elles ont été choisies de façon à ce que les réponses des sujets soient prévisibles, et que les mêmes mots réapparaissent plusieurs fois au cours de chaque expérience. Ainsi, il s'agit principalement de questions très simples, comme par exemple : « De quelle couleur est le ciel ? », « Combien de pattes a un chien ? », ou encore « Combien font 2 + 2 ? ».

5.2.5 Déroulement

Le poste de pilotage du robot était placé dans la chambre sourde (cf. Figure 104). La porte du couloir était grande ouverte. Juste à côté, dans le couloir, se trouvait le haut-parleur diffusant le bruit B, placé contre une véritable machine à café. Sa face avant était recouverte d'un tissu noir afin de dissimuler le sigle « JBL ». Il était ainsi très peu visible, ou pouvait passer pour un élément de la machine à café, également de couleur noire. Dans le couloir était également placées une échelle et une boîte à outils. Dans le plafond, une dalle avait été retirée, et un gros câble aux fils dénudés glissé à travers l'ouverture (cf. Figure 103). Cette mise en scène servait à suggérer que des travaux étaient en cours sur la plateforme. En pratique, bien qu'ils soient tous passés devant l'échelle, aucun des participants n'a fait de remarque à ce sujet.

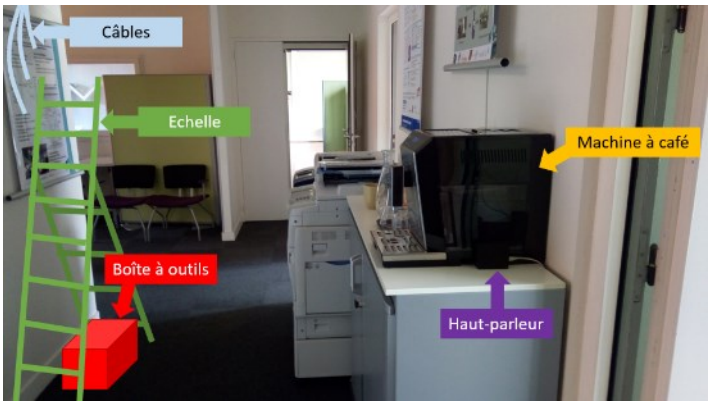


Figure 103 : Photo du couloir avec les éléments de mise en scène

Le second sujet était assis sur une chaise au centre de la plateforme expérimentale, séparé du robot par une petite table basse. La porte de l'appartement domotique était entrouverte, et celle du couloir fermée. Le haut-parleur diffusant le bruit D était placé sur une étagère à l'intérieur de l'appartement domotique et réglé au volume maximal. Du fait des multiples réverbérations du son à l'intérieur de l'appartement, il était difficile de deviner que le son ne provenait pas d'une vraie perceuse.

Pendant le test, nous écoutions le son de la plateforme expérimentale depuis la régie. Les bruits étaient déclenchés de façon régulière, préférentiellement en milieu de question afin qu'ils soient lancés au moment où le sujet R répondait. Pour s'assurer que les sujets ne devinent pas que le son C n'existait que dans le casque du sujet R, le son C n'était jamais déclenché avant plusieurs occurrences du son D. Ainsi, si les sujets mentionnaient la présence du bruit de perceuse, ils comprenaient que celui-ci venait de la pièce du robot. À titre indicatif, la Figure 105 présente une frise chronologique correspondant à une passation de l'expérience.

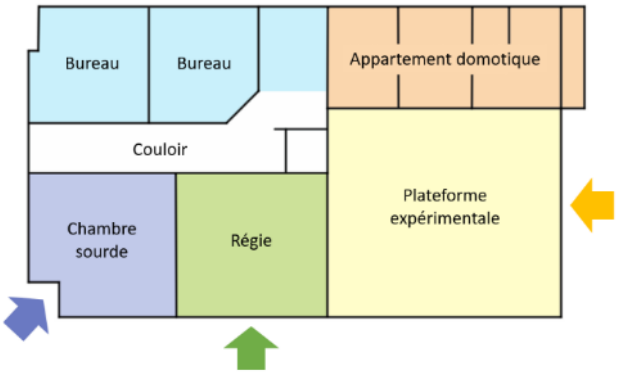


Figure 104 : Plan de Domus (allée de Palestine)

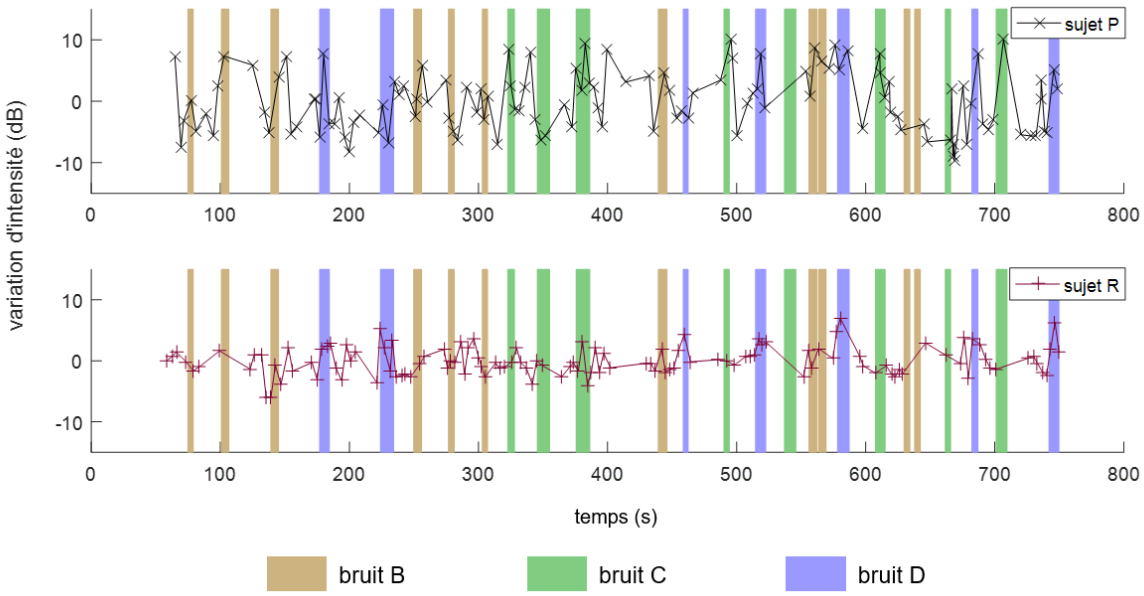


Figure 105 : Exemple de déroulement d'une expérience
Les points de mesure correspondent aux extraits des sujets utilisés pour le modèle (les mots-clés trop rares, apparaissant moins de 50 fois pour tous les sujets ne sont pas pris en compte). Leur abscisse correspond au début de l'extrait. L'ordonnée est la mesure d'intensité maximale de l'extrait, à laquelle a été soustraite la moyenne des mesures.

5.2.6 Débriefing

À la fin de l'expérience, nous réunissions les sujets P et R sur la plateforme expérimentale pour leur demander si l'expérience s'était bien passée, ce qu'ils pensaient de l'interface de téléprésence, et s'ils avaient noté des problèmes pendant la durée du test. Après quelques échanges, nous leur révélions que tester l'interface du robot n'était qu'un prétexte, et leur demandions d'essayer de deviner le but réel de l'expérience. Même à ce stade, la plupart des sujets ne pensaient pas à évoquer les bruits qu'ils avaient entendus : après avoir constaté leur existence, ils les ont tout simplement ignorés pour se concentrer sur la tâche qui leur était confiée. Pourtant, ils ont bien entendu les bruits, et les ont même évoqués à une ou deux occasions pendant l'expérience. Ce n'est que pendant le débriefing, et une fois qu'on leur a dit que les bruits faisaient partie de l'expérience, qu'ils y repensent. Ainsi, une des participantes a mentionné qu'elle s'était inquiétée que les bruits rendent son enregistrement inutilisable ; et un autre a indiqué avoir commencé à tenir son microphone à la main dès le premier bruit pour le rapprocher de sa bouche et être certain d'être audible de son interlocuteur (ses données n'ont malheureusement pas pu être exploitées).

À la fin du débriefing, les sujets avaient la possibilité de revenir sur le formulaire de consentement qu'ils avaient signé avant de commencer l'expérience.

5.2.7 Coulisses de l'expérience

Cette expérience a nécessité plusieurs mises au point techniques, que nous allons décrire dans ce paragraphe. En particulier, il a fallu contrôler précisément l'intensité produite et perçu par le pilote du robot, ainsi que celle des sources de bruit.

5.2.7.1 Matériel et calibration

Le pilote du robot de téléprésence était équipé d'un casque AKG K242, sur lequel était fixé un microphone Sennheiser HSP4. Le signal était transmis par radio, puis numérisé à l'aide d'une carte son UR22MKII. Grâce à l'interface de vidéoconférence, le son était transmis au haut-parleur du robot. Dans l'autre direction, les deux microphones Behringer B5 du robot étaient reliés à une carte son UR33MKI. Le signal stéréo était envoyé grâce à l'interface de vidéoconférence dans le casque du pilote.

Tout d'abord, il a fallu calibrer ce système, afin d'avoir un son de bonne qualité et de conserver l'intensité des signaux transmis. La calibration s'est faite avec l'aide de Coriandre Villain, ingénieur plateformes du Gipsa-lab. Les potentiomètres des cartes sons étaient réglés en buttées pour pouvoir facilement retrouver leurs réglages. Cependant, connaître le réglage exact des haut-parleurs JBL est impossible : leur volume a donc été réglé au maximum. C'est donc uniquement le volume logiciel des périphériques audio qui était réglé.

- Transmission de la voix du pilote

Dans un premier temps, il a fallu fixer le volume du micro utilisé par le pilote pour avoir une bonne dynamique sonore sans saturer. Ensuite, un signal de référence a été enregistré et son intensité mesurée à l'aide d'un sonomètre Lutron SL-4001 : 55,3 dB(A) à 1 mètre. Cet enregistrement a été joué par le haut-parleur du robot, dont le volume a été réglé jusqu'à ce que l'intensité mesurée corresponde à l'intensité de référence.

- Transmission du son binaural

Dans un second temps, un enregistrement de référence de bruit cocktail a été effectué avec les microphones du robot, et son intensité mesurée avec le sonomètre : 56 dB(A) à 1 mètre. L'enregistrement a ensuite été joué dans le casque du pilote. À l'aide d'une oreille artificielle Brüel & Kjar, le volume du casque audio a été réglé jusqu'à ce que l'intensité mesurée soit à nouveau de 56 dB(A). En outre, un test rapide a permis de constaté que la présence du casque n'atténuait pas significativement les sons : la source de bruit de 56 dB(A) mesurée au sonomètre a fourni une mesure de 56 dB(A) avec l'oreille artificielle nue, et de 55,4 dB(A) avec l'oreille artificielle couverte du casque.

- Stabilité des mesures d'intensité

Un test a été effectué pour vérifier si la chaîne audio de Robair (haut-parleur et microphones) est stable lorsque la tablette du robot n'est pas branchée au secteur. Pour ce faire, nous avons lancé un vendredi soir un son test joué en boucle et un enregistrement Audacity. (Par mesure de sécurité, le robot lui, n'était pas allumé, ni branché à la tablette.) L'enregistrement a continué pendant 3h17, jusqu'à extinction de la tablette. Le lundi suivant, nous avons pu étudier l'évolution de l'intensité en comparant deux extraits de dix répétitions du son test, pris au début et à la fin de l'enregistrement. La diminution d'intensité était négligeable : de 69,9 dBPraat au début de la soirée, elle est de 69,3 dBPraat au milieu de la nuit, du fait de la diminution du bruit de fond de la rocade. L'intensité numérique enregistrée aux cours des expériences ne varie donc pas en fonction du niveau de batterie.

5.2.7.2 Annulation d'écho

Durant cette expérience, nous avons utilisé un routeur local, non connecté à internet. L'ordinateur du pilote était relié au routeur par un long câble Ethernet. La tablette du robot et celle de l'expérimentateur étaient reliées au réseau par connexion Wifi. Nous avons mesuré un délai bouche à oreille d'environ 100 ms. Afin de réduire les effets d'écho pour le pilote, le volume de son casque était réduit lorsqu'il parlait, passant immédiatement de 50% à 10% lorsqu'un seuil d'intensité prédéfini était franchi. Lorsque le pilote cessait de parler, le volume de son casque remontait progressivement pendant 200 ms. Cette solution est loin d'être parfaite, mais a l'avantage d'être simple à implémenter. En outre, elle permet d'assurer que le bruit C (injecté directement dans le casque du pilote) soit atténué de la même manière que le bruit D (diffusé dans la pièce robot).

Bien qu'elle en soit inspirée, il ne s'agit pas d'une communication *half-duplex* : au lieu de couper les microphones du robot, on coupe le signal reçu par le pilote. Cela signifie qu'en cas de fluctuations dans la latence de la télécommunication, le pilote peut toujours entendre des échos de sa voix.

5.2.7.3 Gestion des haut-parleurs

Le déclenchement à distance des bruits était contrôlé à l'aide d'une interface très simple, développée par Emeline Le Goff et montrée en Figure 106. Elle était affichée dans un navigateur connecté au serveur ROS du robot. Chaque fois que l'expérimentateur cliquait sur un des trois boutons, un message ROS était envoyé. Les trois appareils connectés au serveur (tablette de l'expérimentateur, ordinateur portable du pilote et tablette du robot) recevaient le message. En fonction du contenu du message (1, 2 ou 3), un bruit était joué par un des trois appareils (1 : JBL du couloir | 2 : casque du pilote | 3 : JBL de l'appartement domotique) à l'aide d'un nœud ROS écrit en Python.

Figure 106 : Interface de déclenchement des bruits
Les numéros 1, 2 et 3 correspondent respectivement aux bruits B, C et D

Notons qu'au bout de quelques minutes d'inactivité, les haut-parleurs JBL s'éteignent. Pour éviter ce problème, nous avons créé un fichier son muet, qui était diffusé en boucle à partir de la mise en place et jusqu'à la fin de l'expérience.

5.3 Prétraitements des données

Dans cette partie, nous décrivons les données recueillies au cours de l'expérience, ainsi que les prétraitements effectués avant l'analyse statistique.

5.3.1 Signaux enregistrés

Plusieurs signaux étaient enregistrés à chaque séance :

1) Les **voix des deux sujets** étaient captées à l'aide de microphones HF, portés au niveau de la bouche. Les récepteurs HF étaient placés dans la chambre sourde et reliés à l'ordinateur portable du pilote à l'aide d'une carte son stéréo UR22MKII. Le signal stéréo était enregistré pour analyse, et la voix du pilote séparée pour être envoyée au haut-parleur du robot.

2) Le **microphone interne** de l'ordinateur portable du pilote servait à enregistrer le bruit dans la pièce du pilote. C'est en particulier grâce à cet enregistrement que les occurrences du bruit B ont été étiquetées.

3) Le **signal de monitoring** du casque porté par le pilote a été enregistré afin de conserver une trace de ce que le pilote entendait. C'est ce signal qui a permis de repérer les bruits C et D. Toutefois, le son de l'enregistrement est un peu différent du son réellement entendu par les sujets : en particulier, les variations de volume liées à l'annulation d'écho ne sont pas notables sur cet enregistrement.

4) L'**enregistrement binaural du robot** était également conservé. Sur cet enregistrement apparaissent uniquement les bruits D, ce qui permet par comparaison avec le signal de monitoring de distinguer les bruits C des bruits D. En pratique, nous avons constaté que le timbre des bruits C était légèrement différent de ceux des bruits D, car la porte était restée grande ouverte, et pas simplement entrouverte, le jour de l'enregistrement des bruits C. Cependant, cette différence de timbre n'a semble-t-il pas été perçue par les sujets, et ne modifie pas significativement l'intensité des bruits C par rapport aux bruits D.

Les enregistrements 1-3) étaient lancés à partir de l'ordinateur du pilote à l'aide d'un script bash et placés dans un dossier portant la date et l'heure de l'expérience. L'enregistrement 4) était lancé sur la tablette du robot à l'aide d'une commande définie dans le programme du robot.

5.3.2 Annotations

Après l'expérience, tous les enregistrements étaient importés dans un fichier Audacity afin de pouvoir être facilement annotés (cf. Figure 107). Dans un premier temps, les occurrences de bruit B, C et D étaient repérées grossièrement à l'aide des numéros 1, 2 et 3. Ensuite, les prototypes de questions et les réponses des sujets étaient découpés soigneusement, en faisant précéder chaque étiquette du numéro de bruit correspondant (0 pour les silences). Notons que le signal de monitoring a tendance à se décaler au cours du temps, car la latence de la vidéoconférence n'est pas constante. Une écoute attentive des enregistrements était donc nécessaire pour déterminer les conditions de bruit de chaque extrait. Les marqueurs étaient ensuite exportés dans un fichier .txt, accompagnés de leur temps de début et de fin.

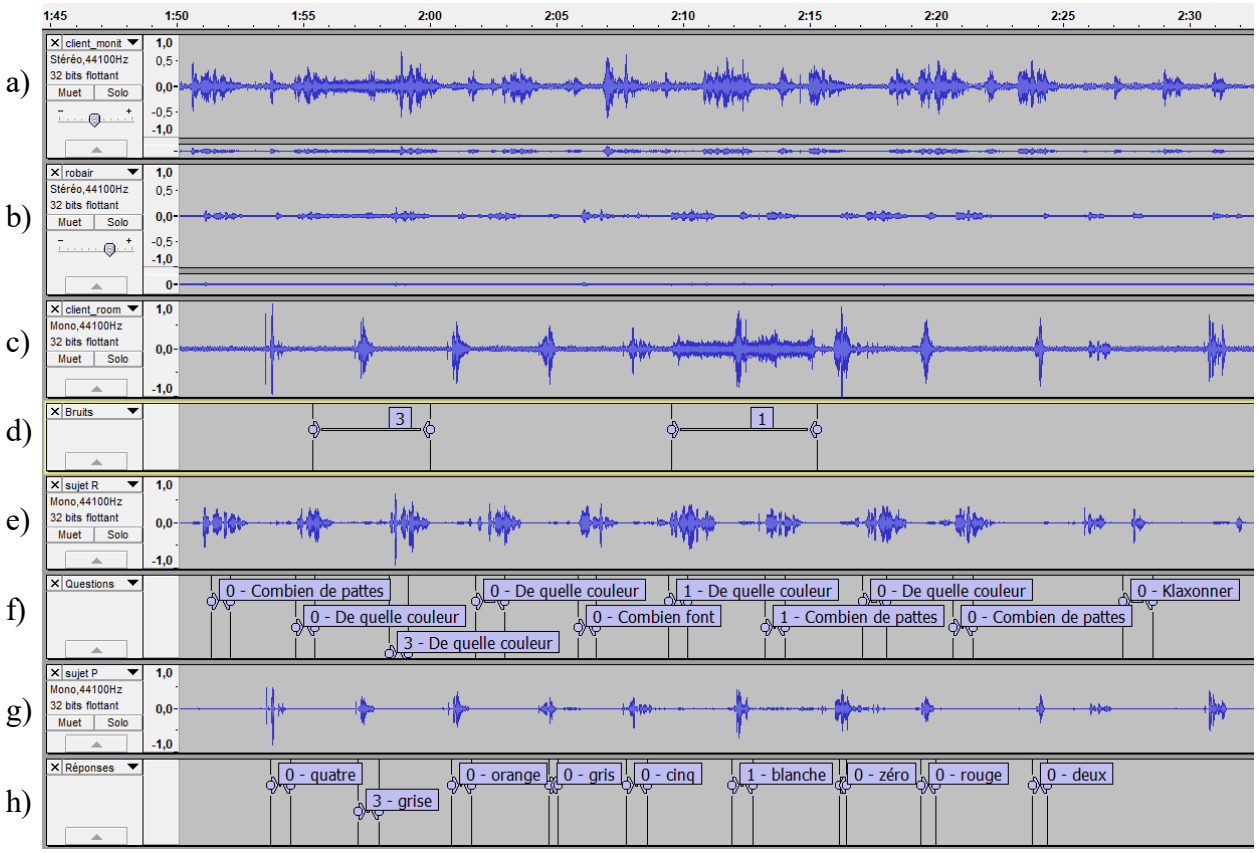


Figure 107 : Interface du logiciel Audacity utilisé pour découper les enregistrements.
a) signal de monitoring – b) enregistrement du robot – c) enregistrement du microphone interne de l'ordinateur du pilote – d) piste de marqueurs des bruits – e) voix du sujet R – f) questions du sujet R – g) voix du sujet P – h) réponses du sujet P

5.3.3 Mesures

Afin de pouvoir mettre en évidence un effet Lombard, nous avons mesuré l'intensité, le pitch et la durée des extraits des sujets P et R.

5.3.3.1 Intensité maximale

La méthode utilisée pour les mesures d'intensité est présentée en Annexe B. Plutôt que d'opter pour une intensité moyenne, nous avons choisi de calculer l'intensité maximale de chaque extrait. En effet, bien que les microphones portés par les sujets se trouvent au niveau de leur bouche, la mesure d'intensité moyenne peut être biaisée lorsqu'il y a un bruit dans la pièce. Au contraire, l'intensité maximale est beaucoup trop élevée pour être impactée sensiblement par les résidus de bruits captés par les microphones.

5.3.3.2 Pitch

Pour les mesures de pitch, nous avons utilisé le logiciel de phonétique Praat, qui permet d'estimer la courbe de pitch. À l'aide d'un script, nous avons généré pour chaque extrait un fichier contenant les mesures de pitch, effectuées toutes les 20 ms. Les courbes de pitch ont ensuite été vérifiées une par une visuellement à l'aide d'un programme Matlab (cf. Figure 108).

Des corrections ont été apportées si nécessaires, le plus souvent, pour supprimer les mesures aberrantes effectuées dans des parties non voisées du signal. Plus rarement, le pitch était mal estimé dans des parties voisées, et a été corrigé à l'aide de l'outil de visualisation de Praat (cf. Figure 109).

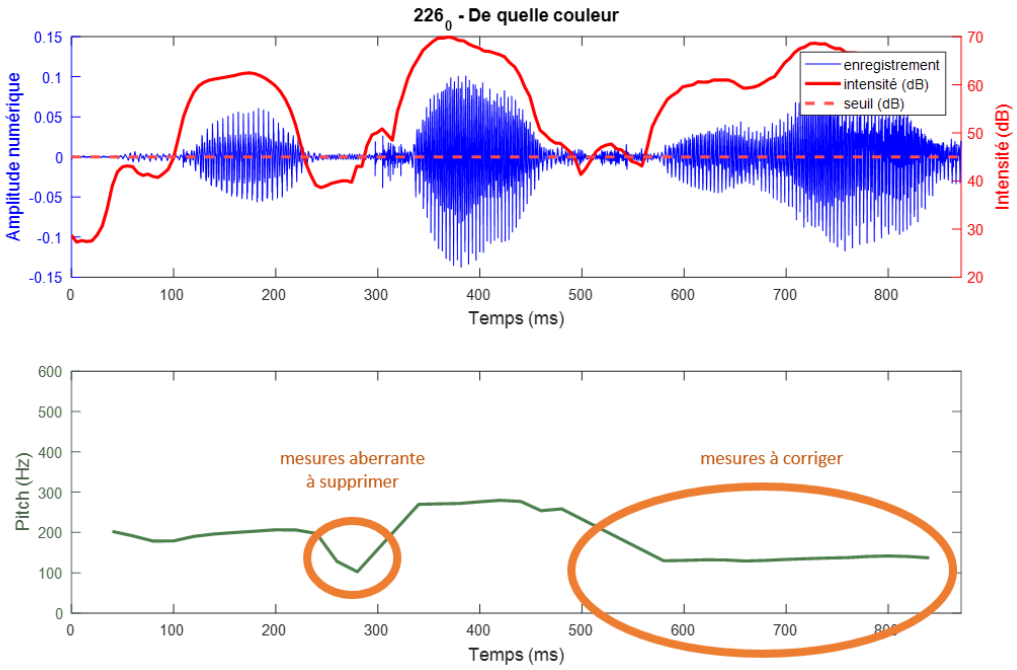


Figure 108 : Visualisation des extraits pour repérer d'éventuelles mesures de pitch aberrantes

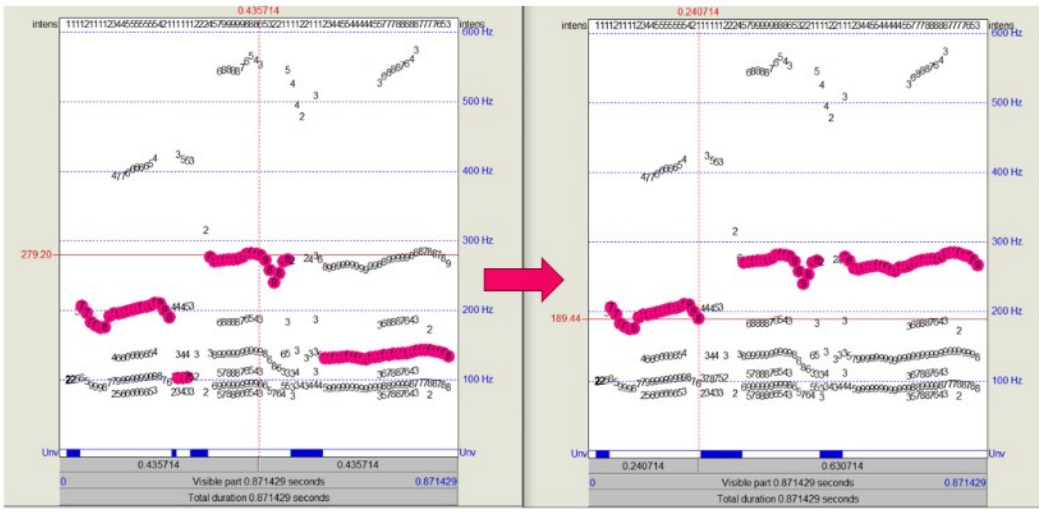


Figure 109 : Correction manuelle des mesures de pitch à l'aide du logiciel Praat

Pour faciliter l'analyse, nous n'avons pas tenu compte de la courbe de pitch complète, mais seulement de sa valeur moyenne, minimale et maximale.

5.3.3.3 Durée

Compte tenu du nombre d'extraits à traiter, nous avons opté pour une mesure automatique de la durée. Deux méthodes différentes ont été utilisées. La première consiste à fixer un seuil sur les mesures d'intensité, et à compter le nombre de trames dont l'intensité est supérieure au seuil. La seconde méthode consiste à compter le nombre de trames pour lesquelles une mesure de pitch est définie : autrement dit, il s'agit de mesurer la durée des segments voisés de l'extrait choisi. Ces deux méthodes ont été testées et validées sur un nombre réduit d'enregistrements.

5.3.4 Quelques statistiques sur les données de l'étude

14 duos de sujets ont participé à l'expérience. La plupart avait le français pour langue maternelle (25/28). Deux autres sujets parlaient couramment le français. Un seul sujet était peu à l'aise avec le français : il a joué le rôle du sujet R.

En moyenne, le nombre de bruits B, C et D joués à chaque séance était respectivement de 10, 7 et 7. Le nombre d'extraits de parole obtenu pour chaque sujet était d'environ 101 (silence) et 11 pour chacune des trois conditions de bruit. Le décompte exact d'extraits tous sujets confondus est visible dans le Tableau 49. En pratique, certains sujets ont donné des réponses non standard, en répondant par exemple que la couleur de la moutarde est « moutarde » au lieu de « jaune ». Nous avons donc choisi de ne pas tenir compte des réponses / questions apparaissant moins de 50 fois dans l'ensemble des résultats.

Tableau 49 : Décompte des extraits pour tous les sujets

Condition	Sujets P	Sujets R
A	1521	1322
B	182	127
C	161	139
D	163	152

5.4 Analyses des résultats

Nous allons à présent analyser les résultats de l'expérience. Ils ont été obtenus à l'aide du logiciel d'analyse statistique R et portent sur les mesures d'intensité, de pitch et de durée. Deux jeux de données séparés sont considérés : d'un côté celles des sujets P, de l'autre celles des sujets R. Pour chaque paramètre étudié, nous présenterons d'abord des résultats globaux, sous forme de boîtes à moustaches. Puis nous modéliserons ces résultats à l'aide d'un modèle linéaire mixte. Pour plus de détails sur les modèles linéaires mixtes, le lecteur est renvoyé au tutoriel de (Winter 2013), ainsi qu'à l'Annexe C.

5.4.1 Intensité

Les résultats les plus remarquables de cette étude concernent l'intensité et sont présentés en Figure 110. On constate que l'intensité maximale varie en fonction des conditions de bruit. Ces variations peuvent être décrites à l'aide du modèle suivant :

$$intensité\ maximale \sim condition\ de\ bruit + (1|mot) + (1|sujet)$$

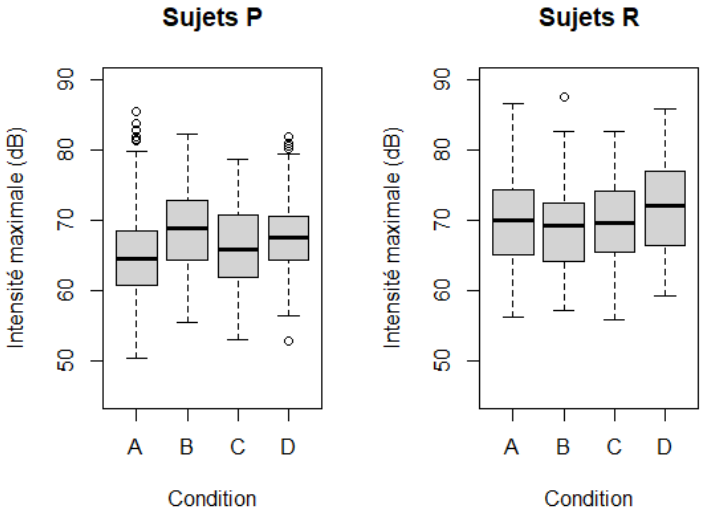
Cette formule signifie que l'intensité maximale est modélisée comme une fonction linéaire de la condition de bruit, mais en tenant compte du fait qu'il existe également une intensité de référence, variable pour chaque mot et chaque sujet. Les résultats de la modélisation sont représentés dans le tableau accompagnant la Figure 110. L'intensité maximale moyenne est estimée uniquement dans la condition A, qui sert de référence. Pour les conditions B, C et D, c'est l'écart à cette valeur de référence qui est indiquée. Le modèle fournit également des erreurs-type, qui représentent l'incertitude sur l'estimation des coefficients du modèle, connus seulement à X dB près. Ces erreurs-types n'ont pas la même signification pour la condition A que pour la condition B : ainsi, si l'intensité moyenne estimée en condition A est de 64,61 ± 1,11 dB pour les sujets P, c'est l'écart entre la condition B et la condition A qui est estimée à 2,98 ± 0,36 dB. L'intensité moyenne estimée par le modèle en condition B serait donc d'environ 64,61 + 2,98 = 67,59, avec une précision de $\sqrt{1{,}11^2 + 0{,}36^2}$ = 1,17 dB.

Le tableau indique également un nombre d'extraits, c'est-à-dire le nombre de données utilisées pour le modèle. Enfin, une mesure de la significativité statistique du modèle a été obtenue, en comparant à l'aide d'une ANOVA les résultats du modèle à ceux du modèle nul :

$$intensité\ maximale \sim 1 + (1|mot) + (1|sujet)$$

Autrement dit, il s'agit de vérifier que le modèle proposé, qui tient compte de la condition de bruit, fournit une meilleure approximation des données qu'un modèle qui se contente de faire la moyenne de toutes les mesures.

Le modèle linéaire proposé semble pertinent, puisque la probabilité d'obtenir les mêmes résultats avec le modèle nul est de l'ordre de 10^{-16} ou moins. Comme prévu, on constate que l'intensité maximale était plus élevée lorsque les sujets entendaient du bruit, c'est-à-dire dans les conditions B, C et D pour les sujets P, et uniquement en condition D pour les sujets R. Pour les sujets P, il existe également des variations entre les trois conditions de bruit, que nous allons détailler ci-dessous.



	Sujets P				Sujets R			
Condition	**A**	**B**	**C**	**D**	**A**	**B**	**C**	**D**
Intensité max (dB)	64,61	+ 2,98	+ 1,19	+ 2,38	69,19	- 0,06	- 0,30	+ 2,24
Écart type (dB)	1,11	0,36	0,37	0,36	1,86	0,21	0,21	0,20
Nombre d'extraits	1263	160	146	150	1125	118	129	143
Significativité	$p < 2.10^{-16}$				$p < 2.10^{-16}$			

Figure 110 : Intensité maximale mesurée

5.4.1.1 Tests de significativité

Pour estimer si les variations entre les conditions de bruit sont significatives, nous avons réduit nos jeux de données à seulement deux conditions, et comparé le modèle linéaire mixte au modèle nul. Les résultats pour les sujets P sont rapportés dans le Tableau 50.

Tableau 50 : Significativité statistique des jeux de données réduits pour les sujets P

	Sujets P					
Conditions étudiées	**{ A , B }**	**{ A , C}**	**{ A , D }**	**{ B , C }**	**{ B , D }**	**{ C , D }**
p	8.3 e^{-16}	1.1 e^{-3}	9.3 e^{-11}	5.5 e^{-5}	8.7 e^{-2}	2.1 e^{-2}
Significativité	***	**	***	***	.	*

La différence d'intensité moyenne entre chaque condition de bruit et la condition de référence est donc significative : dans le pire des cas (conditions { A , C }), la probabilité d'obtenir les mêmes résultats avec le modèle linéaire mixte et le modèle nul est de seulement 1‰. Il y a également un écart significatif entre les mesures obtenues dans les conditions { B , C }. En revanche, l'écart entre les conditions { B , D } et { C , D } est plus faible. À titre de comparaison,

pour les sujets R, il n'y a pas de différence significative entre les conditions A, B et C (cf. Tableau 51).

Tableau 51 : Significativité statistique des jeux de données réduits pour les sujets R

	Sujets R					
Conditions étudiées	**{ A , B }**	**{ A , C}**	**{ A , D }**	**{ B , C }**	**{ B , D }**	**{ C , D }**
p	0.64	0.17	$< 2.2\ e^{-16}$	0.25	$6.7\ e^{-15}$	$< 2.2\ e^{-16}$
Significativité			***		***	***

5.4.1.2 Variabilité inter-sujets

Pour étudier la variabilité inter-sujets, on étudie le modèle linéaire mixte suivant :

$$intensité\ maximale \sim\ bruit + \ (1|mot) + \ (\ bruit|sujet)$$

Cette formule signifie que l'effet aléatoire lié au sujet peut également dépendre de la condition de bruit. Lorsqu'on compare ce modèle au modèle précédent, on n'obtient pas de différence significative (Sujets P : p = 0,69 | Sujets R : p = 0,18). Cela signifie que prendre en compte la variabilité inter-sujets n'apporte pas grand-chose au modèle : les pentes de chaque sujet sont suffisamment proches pour pouvoir être résumées par une moyenne unique. Pour s'en convaincre, regardons en détails les coefficients associés à ce modèle (cf. Figure 111). On constate que tous les sujets suivent de façon plus ou moins marquée les tendances notées au paragraphe précédent : courbe en /\ pour les sujets P et en __/ pour les sujets R.

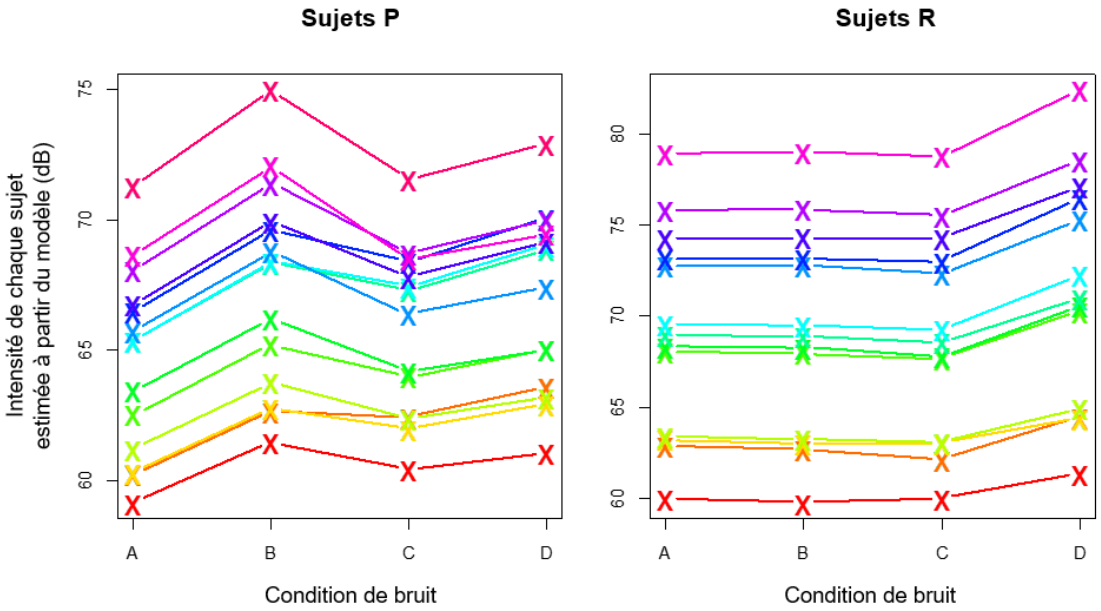


Figure 111 : Modèle linéaire aux effets mixtes, sujet par sujet

Dans la suite, une interprétation de ces résultats est proposée.

5.4.1.3 Condition B et C

On constate que la condition B est celle avec la plus forte augmentation d'intensité : presque 3 dB. C'est le double de l'augmentation constatée en condition C alors que les deux conditions sont similaires : seul le sujet P entend le bruit, qui provient soit d'un haut-parleur situé dans le couloir adjacent (condition B), soit du casque qu'il porte sur la tête (condition C). On s'attendrait donc à observer les mêmes résultats dans les deux conditions, voire une augmentation plus forte en condition C, puisqu'a priori, le sujet P croit que le bruit vient de l'environnement du sujet R, et doit donc parler plus fort pour se faire entendre. En outre, (Garnier et al. 2010) ont montré que l'effet Lombard est plus fort lorsque l'écoute se fait au casque plutôt qu'avec deux haut-parleurs latéraux, car le port du casque atténue le retour auditif que le locuteur a de sa propre voix.

Nos résultats vont donc dans le sens inverse des résultats attendus. Cela peut être simplement dû au fait que les bruits B et C utilisés sont très différents, alors que (Garnier et al. 2010) avaient pris soin d'utiliser le même bruit, et de calibrer soigneusement leur matériel de sorte que l'intensité perçue par les sujets soient la même pour l'écoute au casque et l'écoute aux haut-parleurs. Dans notre cas, le volume des sources de bruit n'a été réglé que de manière approximative, à l'aide d'une mesure à court terme pendant les phases stationnaires des bruits.

5.4.1.4 Condition C et D

Il existe également un écart notable entre les conditions C et D : l'intensité maximale moyenne est plus élevée d'environ 1,2 dB dans le cas D. Or, les sujets P entendent les mêmes bruits dans les deux conditions. La seule différence est qu'en condition D, les sujets R entendent également le bruit et ont tendance à parler plus fort. Cette variation d'intensité chez les sujets P correspond donc sans doute à un effet d'entraînement : les sujets P s'adaptent non seulement au bruit, mais également à l'intensité vocale de leur interlocuteur. Cet effet a déjà été observé par (Szekely et al. 2015) dans le cadre d'une expérience faisant intervenir un sujet naïf, et un expérimentateur qui se force à parler à une intensité de voix donnée : faible, modale ou forte.

En revanche, cet effet n'a pas été constaté chez les sujets R dans les conditions B et C, puisque leur intensité de voix n'a pas augmenté significativement par rapport à la condition de référence A. Cela peut être dû au fait que les sujets R devaient lire des questions, tandis que les sujets P devaient leur répondre. En effet, plusieurs sources montrent que l'effet Lombard est d'autant plus fort que la tâche étudiée est interactive. Ainsi, (Amazi Deborah K., Garber Sharon R. 1982) ont étudié deux groupes de sujets : un groupe devait raconter une histoire à partir d'une série d'images, l'autre devait simplement nommer ces images. Ils ont constaté que l'augmentation d'intensité en présence de bruit était plus importante pour le premier groupe que pour le deuxième. (Junqua et al. 1999) ont également observé que leurs sujets parlaient plus fort lorsqu'ils discutaient avec un système de dialogue que lorsqu'ils devaient simplement lire une série de phrase à ce même système de dialogue. Enfin, (Garnier et al. 2010) se sont intéressées à l'évolution de plusieurs paramètres vocaux en présence de bruit : l'intensité de voix bien sûr, mais également sa fréquence fondamentale, la fréquence du 1er formant, le centre de gravité du spectre et l'ouverture de la bouche. Elles ont constaté que la valeur de tous ces paramètres

vocaux augmentait en présence de bruit, et que l'augmentation était plus importante lorsque les sujets devaient interagir avec un partenaire, que lorsqu'ils devaient parler seuls.

5.4.1.5 Ordres de grandeur des variations

Si les variations d'intensité mesurées sont significatives, elles sont tout de même très faibles : de l'ordre de 1 à 3 dB. À titre de comparaison, le Tableau 52 récapitule les résultats obtenus par quelques études utilisant des bruits de faible intensité. Ainsi, pour des bruits d'environ 55 – 60 dB SPL, les augmentations d'intensité rapportées dans la littérature sont plutôt de l'ordre de 3 à 5 dB, même pour des tâches très peu interactives pendant lesquels les sujets n'ont pas besoin de se faire entendre, puisqu'ils parlent seuls. L'article de (Winkworth Alison L., Davis Pamela J. 1997) rapporte même une augmentation d'environ 12 dB pour un niveau de bruit équivalent au notre.

Plusieurs explications peuvent être avancées pour expliquer ces différences. Tout d'abord, toutes les études n'ont pas la même manière de mesurer l'intensité. Dans notre cas, il s'agit d'intensité numérique maximale, mesurée sur un ensemble fini de « mots ». Dans les autres expériences, il s'agit d'une mesure de pression acoustique moyenne sur plusieurs secondes d'enregistrement, parfois en supprimant certaines portions du signal : par exemple, les enregistrements de rire et d'expiration dans le cas de (Winkworth Alison L., Davis Pamela J. 1997).

Une autre piste d'explication concerne le type de bruit utilisé, qui pourrait avoir une influence sur l'effet Lombard. En particulier, le bruit cocktail utilisé par (Winkworth Alison L., Davis Pamela J. 1997) est peut-être plus dérangeant pour les sujets qu'un bruit blanc filtré, ou un bruit de machine. En effet, un bruit cocktail est obtenu à partir de plusieurs enregistrements de voix humaines, et il est possible en l'écoutant de reconnaître des morceaux de phrase, ou d'identifier des voix.

La durée de l'expérience pourrait également avoir un impact : dans notre cas, les bruits étaient extrêmement brefs, donc les sujets n'avaient que quelques centaines de millisecondes pour s'y adapter. Au contraire, (Marxer et al. 2018) ont prévu par mesure de précaution un temps de chauffe au moment d'enregistrer leur corpus de parole Lombard et standard : pour chaque série de 15 phrases enregistrées dans une condition, les 5 premières étaient mises de côté. Dans les autres expériences citées, le bruit était généralement diffusé pendant toute la durée de l'enregistrement, c'est-à-dire quelques minutes.

Enfin, dans toutes les études sur l'effet Lombard que nous avons consultées, il était évident pour les sujets que les chercheurs s'intéressaient à leur manière de parler dans le bruit, ou du moins que le bruit faisait partie de l'expérience. Au contraire, nos sujets étaient convaincus que les bruits étaient accidentels et n'y ont donc pas particulièrement prêté attention, jusqu'à peut-être même oublier leur existence, puisqu'ils ne pensaient pas spontanément à les mentionner lors du débriefing.

Tableau 52 : Résultats de quelques études sur l'effet Lombard ayant recourt à des bruits d'intensité faible

Etude	Type de bruit	Tâche	Intensité du bruit	Variation d'intensité
(Bottalico 2018)	Bruit de restaurant	Lecture de phrases face à un auditeur	35 dB SPL	Référence
			55 dB SPL	+ 6 dB
			85 dB SPL	+ 22 dB
(Tufts, Frank 2003)	Bruit rose	Lecture à voix haute avec ou sans bouchons d'oreille (mesures ci-contre sans bouchon d'oreille)	60 dB SPL	+ 5 dB SPL
			70 dB SPL	+ 7 dB SPL
			80 dB SPL	+ 10 dB SPL
			90 dB SPL	+ 14 dB SPL
			100 dB SPL	+ 18 dB SPL
(Winkworth Alison L., Davis Pamela J. 1997)	Bruit cocktail	Lecture à voix haute et monologue	55 dB SPL	+ 12 dB SPL
			70 dB SPL	+ 18 dB SPL
(van Heusden et al. 1979)	Bruit blanc filtré pour avoir la même enveloppe spectrale qu'un bruit de conversation	Dialogue spontané	35 dB SPL	+ 0 dB SPL
			45 dB SPL	+ 1,5 dB SPL
			55 dB SPL	+ 4 dB SPL
			65 dB SPL	+ 8 dB SPL
(Korn 1954)	Bruit blanc filtré pour avoir la même enveloppe spectrale qu'un bruit de conversation	Lecture de phrases face à un auditeur	50 dB	+ 2 dB
			60 dB	+ 5 dB
			70 dB	+ 9 dB
			80 dB	+ 13 dB
			90 dB	+ 17 dB

5.4.2 Pitch

Nous avons suivi une démarche similaire, à base de modèles linéaires mixtes, pour étudier le pitch et la durée. D'après nos connaissances a priori sur l'effet Lombard, nous nous attendions dans les conditions de bruit à observer une augmentation du pitch et de la durée, mais moins importante que l'augmentation d'intensité constatée précédemment.

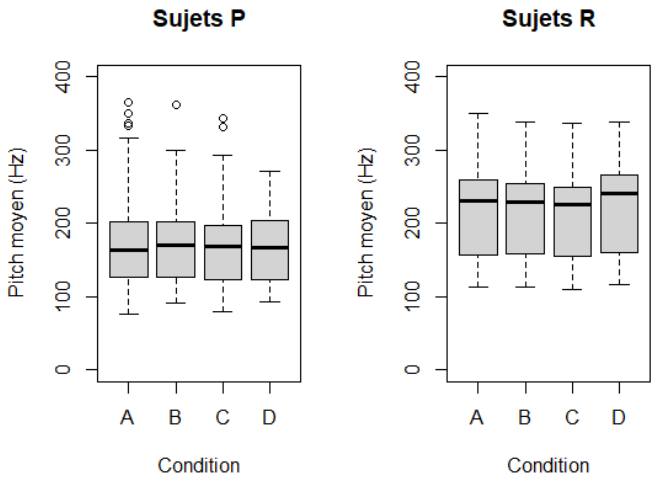
La Figure 112 présente les résultats obtenus pour le pitch moyen. Tout d'abord, on constate qu'en moyenne, le pitch des sujets R est plus élevé. Cette différence s'explique à la fois par la composition des groupes, et par la tâche donnée aux sujets. Ainsi, on compte autant de femmes que d'hommes dans le groupe des sujets P, tandis que le groupe des sujets R compte 8 femmes contre 6 hommes (5 en réalité, car les enregistrements d'un des sujets n'ont pas pu être exploités, car il s'est mis à tenir son micro à la main pour le rapprocher de sa bouche lorsque les bruits ont commencé). En outre, on note que le pitch minimal et le pitch maximal sont plus élevés dans le groupe R que le groupe P : passant respectivement de 100 à 120 Hz, et de 220 à 300 Hz. Bien qu'il soit possible que tous les sujets R aient par hasard une voix plus aiguë que les sujets P, l'hypothèse la plus probable est que cet écart provienne de la tâche donnée aux sujets : en effet, les productions des sujets P sont principalement déclaratives, tandis que celles des sujets R sont toujours interrogatives.

En ce qui concerne les variations de pitch moyen en fonction de la condition de bruit, on constate que dans le cas des sujets P, il n'y a pas de variation significative. En effet, l'erreur-type sur les pentes du modèle ne permet pas de déterminer si la variation de pitch est positive en condition B, C et D. Pour les sujets R en revanche, la p-valeur obtenue indique que les variations sont significatives. En particulier, le pitch moyen est plus élevé en condition D, c'est-à-dire lorsque le bruit est joué dans l'environnement du robot, et légèrement plus faible en condition B/C, c'est-à-dire lorsque le bruit est présent, mais inaudible pour les sujets R. Les tests de significativité sur les jeux de données réduits sont présentés au Tableau 53.

Tableau 53 : Significativité statistique des jeux de données réduits pour les sujets R (pitch moyen

	Sujets R					
Conditions étudiées	**{ A , B }**	**{ A , C}**	**{ A , D }**	**{ B , C }**	**{ B , D }**	**{ C , D }**
p	1.10^{-2}	9.10^{-4}	6.10^{-6}	4.10^{-1}	5.10^{-6}	2.10^{-8}
Significativité	*	***	***		***	***

Les résultats obtenus avec le pitch minimal et maximal sont similaires.



	Sujets P				**Sujets R**			
Condition	**A**	**B**	**C**	**D**	**A**	**B**	**C**	**D**
Pitch moyen (Hz)	166	+ 2	+ 2	+ 3	218	- 3	- 4	+ 6
Erreur type (Hz)	11	2	2	2	16	1	1	1
Nombre d'extraits	1265	158	146	150	1225	118	129	143
Significativité statistique	$p = 0{,}52$				$p = 3.10^{-8}$			

Figure 112 : Fréquence fondamentale moyenne mesurée

Examinons à présent pourquoi nous n'avons pas pu constater d'augmentation significative de pitch pour les sujets P. Le Tableau 54 présente une estimation de la part de la variance expliquée par les différents effets aléatoires pris en compte dans le modèle. Tout d'abord, on constate que la variance totale est plus faible pour les sujets P que pour les sujets R : elle s'élève respectivement autour de 2300 contre 3300. Cette différence est une conséquence de l'augmentation de pitch constatée précédemment pour le groupe R. Ensuite, on remarque que la part de variance expliquée par le modèle varie en fonction du groupe de sujets. Pour les sujets P, environ 65 % de la variance est attribuée à la variable « sujet », et 5 % à la variable « mot ». Les 30 % restant ne sont pas expliqués par le modèle. Pour les sujets R en revanche, 95 % de la variance est expliquée par le modèle, dont 94 % proviennent de la variable « sujet ». Autrement dit, on a un modèle très pertinent dans le cas des sujets R, et un modèle assez incomplet dans le cas des sujets P, qui ignore une ou plusieurs causes de variabilité. En particulier, on sait que le ton des sujets P était beaucoup moins régulier que celui des sujets R : souvent neutre, mais parfois hésitant ou rieur. Les productions des sujets R au contraire étaient lues, beaucoup plus longues et standardisées. Il manque donc certainement une variable au modèle, permettant de décrire le socio-affect exprimé par les sujets P. Notre hypothèse est que cette variabilité non expliquée masque l'effet fixe recherché, c'est-à-dire une augmentation d'intensité liée au bruit.

Tableau 54 : Estimation de la variance expliquée par les effets aléatoires du modèle

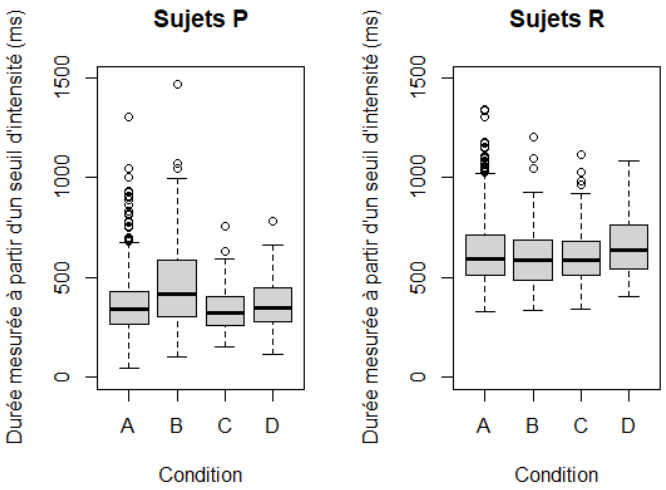
	Sujets P		**Sujets R**	
Variable	**Variance expliquée**	**Erreur type**	**Variance expliquée**	**Erreur type**
sujet	1548.9	10.6	3360.9	58.0
mot	111.7	39.4	3.8	2.0
résidu	725.7	27.0	208.1	14.4

5.4.3 Durée

Pour finir, nous allons nous intéresser à la durée des extraits enregistrés. Deux méthodes ont été envisagées pour la mesurer : une basée sur les courbes d'intensité, l'autre sur les courbes de pitch. Nous allons voir qu'elles fournissent des résultats différents.

Sans surprise, les extraits des sujets R durent plus longtemps que ceux des sujets P, puisqu'il s'agit de débuts de questions, et non de mots uniques.

La Figure 113 présente les résultats obtenus dans le cas où la durée est mesurée en considérant uniquement les segments d'intensité supérieure à 40 dB. Pour les sujets P, on constate une nette augmentation de la durée en condition B : elle représente environ 30 % de la durée de référence. Si l'erreur type ne permet pas de comparer les conditions C et D à la condition A, on note tout de même une augmentation de la durée entre la condition C et la condition D. Pour les sujets R, on n'observe pas de variation significative entre les conditions A, B et C. En revanche, on note une légère augmentation en condition D.



	Sujets P				Sujets R			
Condition	**A**	**B**	**C**	**D**	**A**	**B**	**C**	**D**
Durée (ms)	368	+ 103	- 12	+ 13	620	0	- 7	+ 30
Erreur type (ms)	23	10	10	10	37	10	10	9
Nombre d'extraits	1265	158	146	150	1225	118	129	143
Significativité statistique	$p < 2{,}2\,.\,10^{-16}$				$p = 7{,}1\,.\,10^{-3}$			

Figure 113 : Durée mesurée à partir d'un seuil d'intensité

Ces résultats sont confirmés par les analyses des jeux de données réduits (cf. Tableau 55 et Tableau 56).

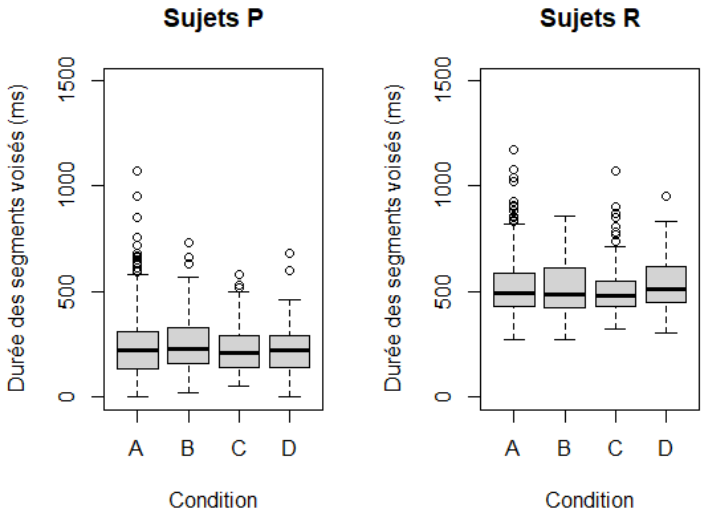
Tableau 55 : Significativité statistique des jeux de données réduits pour les sujets P (durée 1)

	Sujets P					
Conditions étudiées	**{ A , B }**	**{ A , C}**	**{ A , D }**	**{ B , C }**	**{ B , D }**	**{ C , D }**
p	$< 2{,}2\,.\,10^{-16}$	0,2	0,2	$6.\,10^{-9}$	$2.\,10^{-7}$	$8.\,10^{-3}$
Significativité	***			***	***	**

Tableau 56 : Significativité statistique des jeux de données réduits pour les sujets R (durée 1)

	Sujets R					
Conditions étudiées	**{ A , B }**	**{ A , C}**	**{ A , D }**	**{ B , C }**	**{ B , D }**	**{ C , D }**
p	0,9	0,5	9.10^{-4}	0,7	2.10^{-2}	2.10^{-3}
Significativité			***		**	**

Cette première mesure de la durée confirme donc les résultats obtenus pour l'intensité maximale. En revanche, lorsqu'on s'intéresse à la durée des segments voisés, on n'observe aucune variation significative, quel que soit le groupe de sujets ou la condition de bruit considérée (cf. Figure 114). Pour les sujets P, on note bien une augmentation de la durée en condition B, mais trop faible par rapport aux erreurs-type.



	Sujets P				Sujets R			
Condition	**A**	**B**	**C**	**D**	**A**	**B**	**C**	**D**
Durée (ms)	207	+ 17	0	+ 1	506	+1	+ 2	+ 4
Erreur type (ms)	22	8	8	8	36	8	8	7
Nombre d'extraits	1265	158	146	150	1225	118	129	143
Significativité statistique	$p = 0.6$				$p = 0.9$			

Figure 114 : Modèle linéaire mixte associé à la durée mesurée à partir des segments voisés

Il y a deux façons d'interpréter ce résultat : soit il n'y a effectivement pas d'augmentation significative de la durée des sons voisés lorsque les sujets entendent du bruit, et l'augmentation de durée notée précédemment ne concerne que les consonnes non voisées ; soit, la variabilité des mesures est trop importante pour pouvoir conclure.

La seconde hypothèse nous semble plus probable : en effet, on sait que la mesure automatique de pitch n'est pas aussi précise et robuste qu'une mesure à la main. Ainsi, les frontières entre partie voisée et non voisée ont pu être mal estimées. En particulier, nous avons constaté que le pitch des mots « verts » et « quatre » était particulièrement difficile à estimer lorsque les sujets les prononçaient rapidement et d'une voix grave. À titre d'exemple, la Figure 115 montre la première syllabe du mot « quatre », prononcé par deux sujets différents : à gauche, s'il est possible de repérer à la main des impulsions glottiques, les pseudo-périodes sont trop irrégulières pour pouvoir être estimées avec certitude de façon automatique avec Praat ; à droite, le nombre de pseudo-périodes est plus élevé, car la voix est plus aiguë et le mot plus long, ce

qui permet de calculer automatiquement le pitch de façon fiable. Cependant, réaliser des mesures de pitch entièrement à la main serait extrêmement fastidieux. Il n'est pas non plus certain que cela suffise à mettre à évidence des variations intéressantes, en particulier pour les sujets P, puisque nous avons constaté qu'ils n'étaient pas constants dans leur manière de parler. Par manque de temps et face au risque de ne pas obtenir plus de résultats, nous avons donc choisi de nous contenter des mesures automatisées.

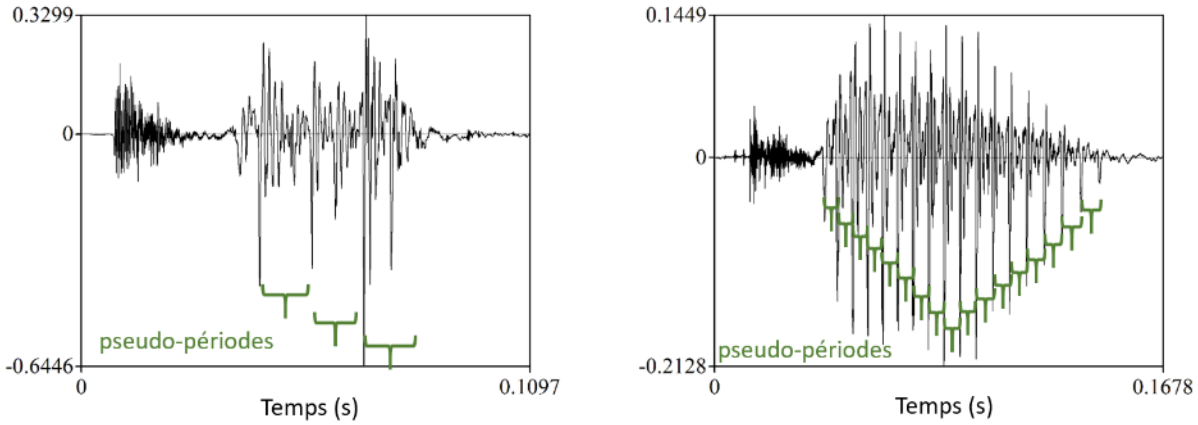


Figure 115 : Première syllabe du mot « quatre » accompagnée de son spectrogramme
A gauche : voix d'homme (environ 80 Hz) | A droite : voix de femme (environ 160 Hz)

5.5 Conclusion

Nous avons conçu une nouvelle expérience, afin de déterminer si l'effet Lombard pouvait altérer la portée vocale en robotique de téléprésence. En effet, une parole Lombard produite uniquement par le pilote du robot pourrait être mal interprété par ses interlocuteurs distants, et confondue avec des variations socio-affectives.

Pour répondre à cette question, nous avons choisi une mise en scène la plus réaliste possible : les bruits choisis sont donc des bruits du quotidien, courts et de faible intensité. De plus, nous nous sommes assuré que les sujets n'y prêtent pas particulièrement attention, en faisant passer ces bruits pour des bruits accidentels, et totalement indépendants de notre volonté. L'avantage de cette démarche est qu'elle permet d'extrapoler les résultats observés en laboratoire à la réalité des usages. Son inconvénient est que les données recueillies sont plus clairsemées et variables que celles qu'on aurait obtenues en découpant simplement l'expérience en quatre sessions A, B, C, D, une pour chaque condition de bruit.

Nous avons pu observer un certains nombres de tendances, concordantes avec la littérature. Ainsi, l'intensité des sujets augmente en présence de bruit ; dans le cas des sujets face au robot, on a même pu constater une légère augmentation de la fréquence fondamentale. De plus, nous avons constaté que le pilote du robot a tendance à parler plus fort, y compris lorsque le bruit n'est pas audible par son interlocuteur, et ne nuit donc pas à l'intelligibilité. Notons également que ces variations d'intensité sont plus faibles que prévues, de l'ordre de 1 à 3 dB seulement. À l'écoute des enregistrements seuls, il est impossible de deviner si les variations d'intensité audibles entre chaque enregistrement sont dues à l'effet Lombard, ou à des variations socio-affectives.

Par ailleurs, il semble exister un effet d'entraînement pour les sujets qui pilotent le robot : ils parlent légèrement plus fort dans la condition où leur interlocuteur parle également plus fort. Il serait intéressant de confirmer ces observations dans le cadre d'expériences plus classiques sur l'effet Lombard, et si cet effet d'entraînement existe bel et bien, de déterminer sa nature : s'agit-il d'un simple réflexe, ou d'une stratégie plus élaborée ?

Chapitre 6 :
APORIA, UN SPECTACLE ARTS-SCIENCES

Dans ce dernier chapitre, nous aborderons le toucher social sous un angle différent, à travers notre participation au spectacle Aporia. Si notre apport est plutôt d'ordre technique que scientifique, il s'agit bien d'une mise en application directe des principes méthodologiques développés au chapitre 3 : pluridisciplinarité, expérimentation en conditions « écologiques » et développement en boucles agiles. De plus, le spectacle aborde d'un point de vue artistique la notion d'altérité et de lien social, la relation intrinsèque entre le corps physique et le corps social, rejoignant ainsi le cadre général dans lequel s'inscrivent nos travaux de recherche.

Aporia est une adaptation de la pièce de théâtre *Combat de nègre et de chiens* de Bernard-Marie Koltès par l'artiste plasticien Alain Quercia. Au cours du spectacle, un acteur unique modifie son toucher social, pour incarner à tour de rôle les différents personnages de la pièce : trois hommes et une femme. Sa performance est augmentée par des technologies numériques, développées en co-construction entre l'équipe artistique et l'équipe scientifique. Nous avons conçu pour cela un système qui permet à l'artiste de transformer en direct la hauteur de sa voix.

Après avoir présenté un rapide historique du projet, nous décrirons en détails l'algorithme utilisé pour réaliser les transformations de voix. Puis, nous raconterons comment l'outil permettant de passer d'une voix à l'autre a été développé, au fil de plusieurs allés-retours entre le Living Lab Domus, le fablab universitaire FabMSTIC, et les scènes de théâtre du Prunier Sauvage et de l'Est.

6.1 Genèse de l'œuvre

Commençons par présenter rapidement la genèse du projet.

Aporia, c'est d'abord une sculpture d'Alain Quercia, inspirée par la lecture de la pièce de Bernard-Marie Koltès : *Combat de nègre et de chiens*. La sculpture représente le combat de deux monstres, deux chimères mi-chiens, mi-hommes. Elles font écho aux personnages de Horn et Cal, deux européens blancs, respectivement chef de chantier et ingénieur sur un site de construction en Afrique. Ils sont confrontés à Alboury, un mystérieux homme noir qui vient réclamer le corps de son frère disparu, et qui ne cédera devant aucune menace ou corruption pour le récupérer. Face à cet « autre » radical, l'égoïsme et la violence des deux hommes sont révélés jusqu'à l'autodestruction : le premier perd sa femme, le second la vie.

Figure 116 : Les Chimères d'Alain Quercia

L'objectif d'Alain Quercia est de créer une œuvre capable d'estomper la frontière entre sculpture, théâtre, public et artiste. En s'adressant au LIG par l'intermédiaire de Nicolas Balacheff, sa première idée était d'animer les Chimères, afin qu'elles réagissent aux spectateurs et conservent une trace de leurs interactions. En août et septembre 2016, il réalise une première résidence à Domus : avec l'aide de Jérôme Maisonnasse, à l'époque responsable du fabMSTIC, il anime le regard des chimères grâce à une projection lumineuse en 3D, afin qu'elles puissent suivre les spectateurs des yeux. C'est à cette occasion que Véronique Aubergé lui propose d'utiliser notre algorithme de conversion de voix en temps-réel. Par la suite, cet outil est développé et testé au cours de plusieurs résidences, à Domus, et sur deux scènes de théâtre : au Prunier Sauvage en juin 2017, et à l'Est en juillet et octobre 2018.

6.2 Algorithme de conversion de voix en temps-réel

Dans cette seconde section, nous présenterons l'algorithme de conversion de voix utilisé dans le spectacle Aporia. Après avoir décrit son historique et son principe général, nous entrerons dans des détails de l'algorithme et de son implémentation. Enfin, nous nous intéresserons au son des voix modifiées par cet algorithme.

6.2.1 Origine de l'algorithme

Initialement, cet algorithme a été développé avec Gang Feng au cours d'un stage de deuxième année d'école d'ingénieurs (Davat 2015). Il s'agit d'une adaptation au temps-réel de l'algorithme de time-stretching conçu par Feng, qui permet de convertir la voix d'une personne pour la rendre plus aigüe ou plus grave, sans pour autant changer ses caractéristiques prosodiques. En particulier, cet algorithme a été utilisé pour des expériences en Magicien d'Oz menées au cours de la thèse de (Sasa 2018). Au cours de ces expériences, des personnes âgées interagissaient avec un robot émettant des micro-expressions vocales (bruits de bouches, rires, interjections…). Le robot, présenté comme autonome, était en réalité piloté à distance par des expérimentateurs, qui sélectionnaient au fur et à mesure les sons et mouvements produits par le robot parmi une liste prédéfinie. Chaque fois que le robot devait « dire » quelque chose de nouveau, il fallait d'abord enregistrer la voix d'un des « magiciens », puis la transformer, avant de pouvoir enfin l'ajouter à la liste des sons possibles. Ces traitements étaient difficiles à concilier avec une interaction spontanée, d'où l'intérêt d'une implémentation en temps-réel.

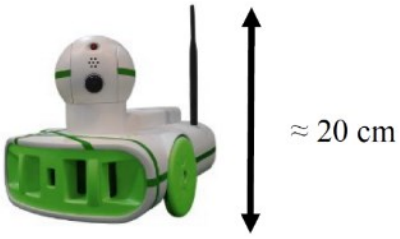


Figure 117 : Le robot Emox, développé par l'entreprise Awabot et utilisé au cours de la thèse de Yuko Sasa

Une première implémentation en temps-réel avait déjà été proposée par (Prablanc 2014). Cependant, elle n'avait pas pu aboutir entièrement, en dépit de son travail sérieux et du temps investi : il subsistait des *glitchs* audio, que Prablanc attribuait à des erreurs dans la bibliothèque de gestion des périphériques audio. De plus, cette implémentation ayant été faite directement en C, Prablanc avait beaucoup de difficultés à visualiser les signaux manipulés, ce qui a très certainement ralenti son temps de développement. En nous appuyant sur ses travaux antérieurs et sur l'environnement de calcul scientifique Matlab, nous avons pu réaliser des transformations de voix en temps-réel d'excellente qualité.

Dans les paragraphes suivant, nous expliquerons le principe de cet algorithme, et de son adaptation au temps-réel, en nous attardant sur les étapes les plus critiques.

6.2.2 Principe du time-stretching

Une technique très ancienne pour modifier des enregistrements de voix consistent à utiliser une vitesse de lecture différente de la vitesse d'enregistrement. Par exemple, en jouant un son deux fois plus vite que prévu, on multiplie par deux toutes ses fréquences, y compris sa fréquence fondamentale : le son paraît donc plus aigu. Cette technique a notamment été utilisée en 1958 par Ross Bagdasarian pour créer les voix des très aigues de « Alvin et les Chipmunks », groupe fictif constitué de trois écureuils et leur père adoptif. Cependant, le son joué deux fois plus vite est également deux fois plus court : pour que les Chipmunks chantent à un rythme donné, Bagdasarian devait donc chanter très lentement, afin que le rythme de sa voix accéléré ne soit

pas trop rapide. Cette technique d'enregistrement ne convient évidemment pas pour une application temps-réel : pour que la voix modifiée soit jouée en simultané, sa durée doit être la même que celle de la voix initiale.

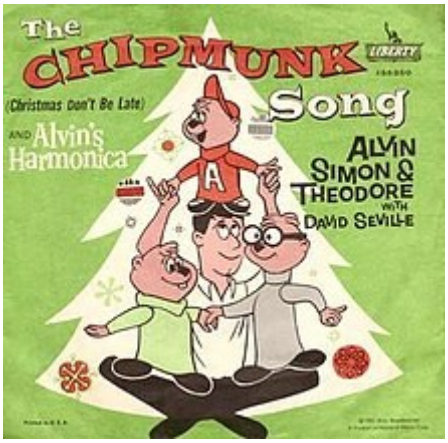


Figure 118 : Pochette du 1er album des Chipmunks (réédition de 1961)
Très populaire aux États-Unis, le groupe apparaît dans plusieurs films et séries d'animations.

Le but du time-stretching est donc de parvenir à conserver ce rythme, afin que la voix modifiée dure le même temps que la voix initiale. Seules les fréquences doivent varier. Pour ce faire, il est nécessaire de modifier artificiellement la durée du signal. Ainsi, pour une obtenir une voix plus aigüe, on commencera par créer un signal de durée allongée, en ajoutant régulièrement des morceaux de signal, avant de jouer ce signal en accéléré. Au contraire, pour obtenir une voix plus grave, on coupera des morceaux du signal afin de raccourcir sa durée, pour ensuite jouer ce signal au ralenti. Ce principe est illustré de façon simplifiée en Figure 119.

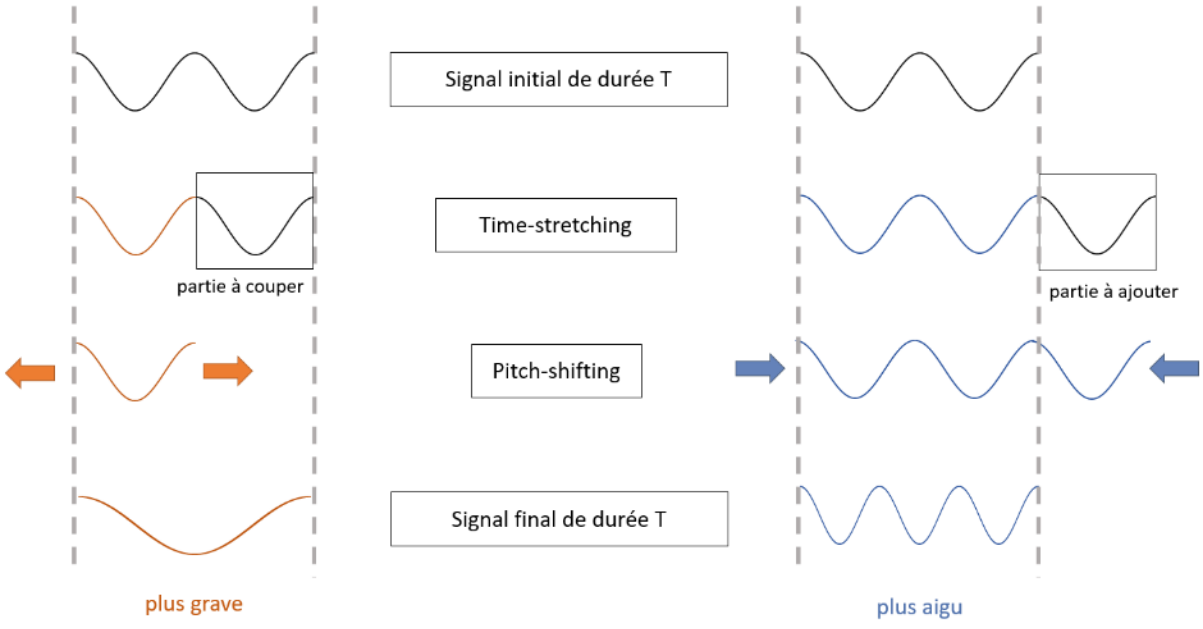


Figure 119 : Principe de la modification de pitch par time-stretching

6.2.3 Algorithme de time-stretching

En pratique, un signal de parole est beaucoup plus complexe qu'une simple sinusoïde : le cœur de l'algorithme de time-stretching consiste donc en une méthode astucieuse pour ajouter / couper des morceaux du signal. Cette méthode est illustrée en Figure 120 pour un rapport de transformation de 1,6 ; c'est-à-dire que la durée du signal est multipliée par 1,6. Ici, la fréquence d'échantillonnage est de 44,1 kHz, donc 500 échantillons correspondent à environ 11 ms.

Tout d'abord, le signal est découpé en trames redondantes, de taille judicieusement choisie : chaque trame doit être suffisamment longue pour pouvoir calculer une intercorrélation, mais suffisamment courte pour que le signal soit relativement stationnaire à l'intérieur de chaque trame. Ces trames sont ensuite décalées dans le temps afin d'allonger / réduire la durée du signal. Localement, leur position précise est calculée de manière à maximiser l'intercorrélation entre deux trames successives. Ceci permet d'assurer que la différence de phase entre les trames soit nulle. Enfin, la jonction entre les trames est effectuée à l'aide de la technique Overlap Add (OLA). Il s'agit d'un simple fondu audio : le son de la première trame est remplacé progressivement par celui de la deuxième, de sorte que la transition entre les deux est imperceptible.

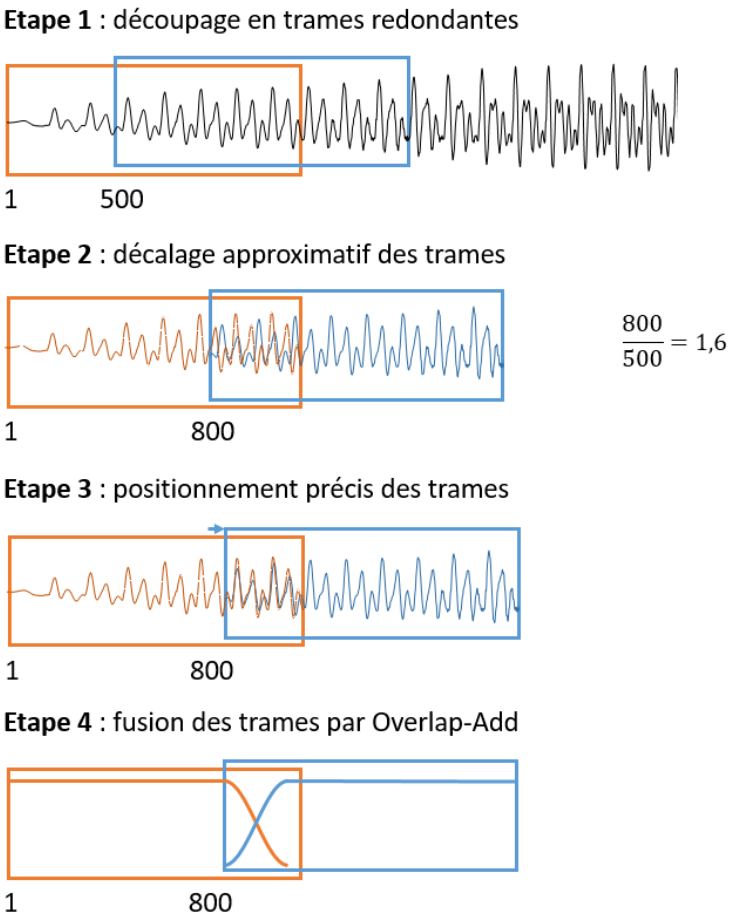


Figure 120 : Illustration du time-stretching pour un rapport de transformation de 1,6

6.2.4 Rééchantillonnage

Une fois le signal de longueur souhaitée obtenu, il faut changer sa vitesse pour le ramener à sa durée initiale. Pour ce faire, on va jouer sur sa fréquence d'échantillonnage.

En effet, contrairement à ce que peuvent laisser croire les figures précédentes, le signal manipulé n'est pas continu, mais échantillonné à une certaine fréquence Fe qui est la fréquence d'enregistrement. Pour pouvoir entendre le signal tel qu'il a été enregistré, il faut le jouer à la même une fréquence Fe. Si on augmente la fréquence de lecture, le signal sera joué plus vite ; et si on la diminue, il sera joué plus lentement. Ainsi, notre signal de longueur multipliée par 1,6 devrait être joué à 44,1 × 1,6 = 70,6 kHz pour durer le même temps que le signal initial.

En pratique, les appareils audio sont conçus et optimisés pour fonctionner à des fréquences d'échantillonnage précises : ce n'est donc pas une bonne idée d'utiliser des fréquences d'échantillonnage « exotiques » telles que 70,6 kHz. Pour contourner le problème, on procède donc à un rééchantillonnage, dont le principe est illustré en Figure 121 : il s'agit de faire comme si le signal avait été enregistré à 44,1 / 1,6 = 27,6 kHz, pour ensuite le jouer à 44,1 kHz.

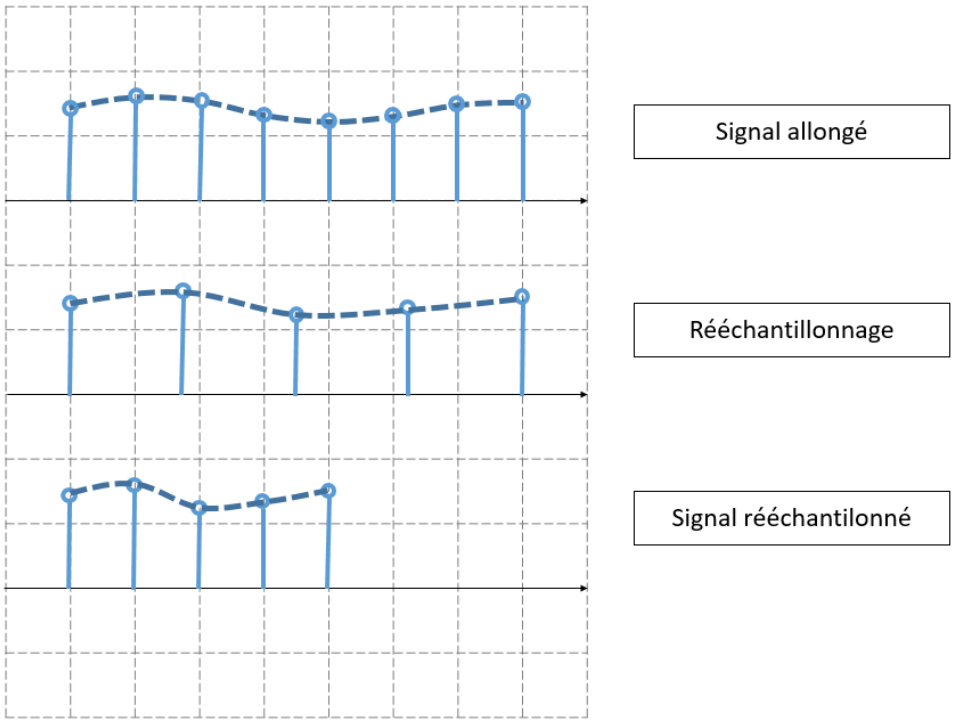


Figure 121 : Principe du changement de vitesse par rééchantillonnage

6.2.5 Principe des traitements en temps-réel

L'algorithme de G. Feng fonctionne trame par trame, c'est-à-dire en considérant le signal uniquement par petits intervalles de temps. Il est donc particulièrement adapté à une implémentation en temps-réel, qui repose sur l'utilisation de buffers[27] : un buffer d'entrée, qui enregistre le signal d'un microphone, et un buffer de sortie, qui transmet ses données à un haut-parleur (cf. Figure 122). Le buffer d'entrée contient T millisecondes de signal audio. Toutes les T millisecondes, les données du buffer sont effacées et remplacées par les T millisecondes suivantes du signal. L'enjeu est donc de réussir à traiter les données du buffer d'entrée dans le temps imparti T, pour pouvoir les placer dans le buffer de sortie.

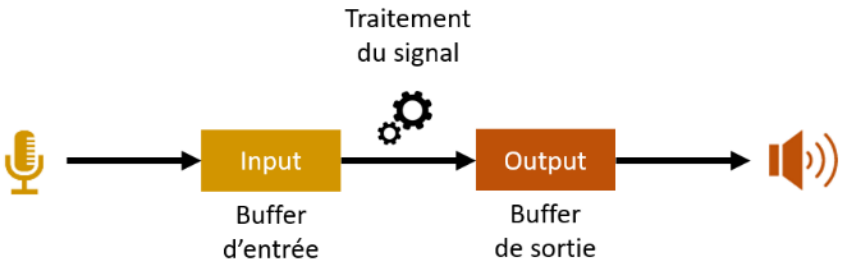


Figure 122 : Principe du traitement d'un signal audio en temps-réel

Étant donné la simplicité de l'algorithme, il n'est pas difficile de respecter cette contrainte de temps. Cependant, le fait de travailler trame par trame peut poser quelques soucis au niveau de l'implémentation que nous allons aborder dans le paragraphe suivant.

6.2.6 Principaux écueils à l'implémentation

6.2.6.1 Conception du temps

L'algorithme de time-stretching repose sur l'utilisation de deux pointeurs temporels : pointeur1 sur le signal initial, et pointeur2 sur le signal allongé/raccourcis. Tout ce qui se trouve avant le pointeur a déjà été traité par l'algorithme, et ne sera plus modifié par la suite.

Pour la version hors temps-réel de l'algorithme, les deux pointeurs sont incrémentés régulièrement. À chaque étape, on lit une trame du signal initial à partir de l'instant pointé par pointeur1, et on copie une trame dans le signal allongé/raccourcis autour de l'instant pointé par pointeur2. Le programme se termine lorsqu'on ne peut plus lire de nouvelle trame.

Dans le cas d'une implémentation en temps-réel, on ne dispose pas du signal en entier. Il est interdit d'enregistrer chaque nouveau buffer à la suite du précédent, car on ne dispose pas d'une mémoire infinie, et le programme est censé pouvoir tourner pendant un temps infini. Au lieu de décaler les pointeurs vers la droite, c'est le signal lui-même qui se décale progressivement vers la gauche. À chaque instant, on n'a accès qu'à une fenêtre du signal. Il faut donc s'assurer que les pointeurs et les morceaux de signaux traités restent bien à l'intérieur de cette fenêtre.

[27] Zone de mémoire qui enregistre temporairement des données.

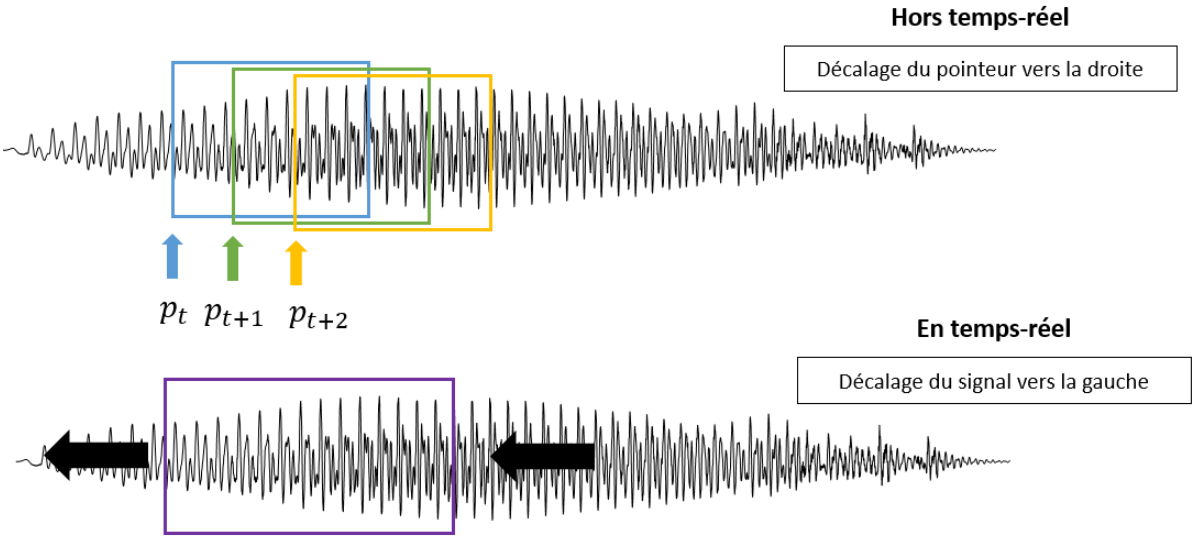


Figure 123 : Différentes manières de concevoir le temps en fonction du type d'implémentation

6.2.6.2 Gestion des effets de bord

Dans la version initiale de l'algorithme, le signal était d'abord allongé / raccourci en entier, avant d'être rééchantillonné pour changer son pitch. Pour l'implémentation en temps-réel, il a fallu effectuer ce rééchantillonnage trame par trame. C'est là que réside la principale difficulté. En effet, le rééchantillonnage implique un filtrage, et engendre donc des effets de bord. Ces effets de bord apparaissent parce que pour calculer un échantillon, le filtre a besoin des N échantillons qui l'entourent, N correspondant à la taille du filtre. Pour pouvoir calculer les échantillons des bords, on est donc obligé d'inventer les échantillons manquants : par exemple, on suppose que ces échantillons ont une valeur de 0. Il y a donc une erreur dans le calcul, mise en évidence en Figure 124.

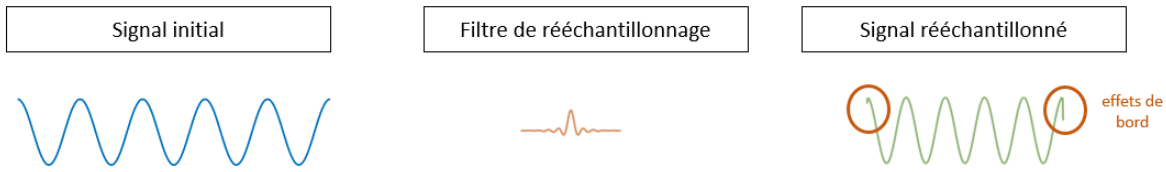


Figure 124 : Mise en évidence des effets de bord provoqué par un rééchantillonnage

Pour corriger ces effets de bord, il est nécessaire d'introduire une marge d'erreur. Ainsi, pour rééchantillonner un signal de 500 échantillons avec un filtre de 121 échantillons, on considérera une fenêtre un peu plus large, de 560 échantillons. Les échantillons des marges seront ensuite coupés pour conserver uniquement le centre de la fenêtre. Notons qu'il n'est pas nécessaire de prendre une marge égale au nombre d'échantillons manquants, car les erreurs de calcul deviennent négligeables dès qu'on s'éloigne des extrémités du signal.

6.2.7 Choix des transformations

Le pitch des voix du robot Emox ou des Chipmunks a été très fortement modifié. En effet, il fallait que ces voix soient très aigues pour qu'elles soient cohérentes avec la petite taille de leur propriétaire. Cette transformation extrême avait également l'intérêt de rendre la voix inhumaine. Au contraire, pour le spectacle Aporia, nous avons choisi des modifications de pitch relativement faibles, inférieures à 20%, pour obtenir des voix humaines. Les valeurs exactes pour chaque personnage sont présentées dans le Tableau 57.

Tableau 57 : Réglages du pitch pour chaque personnage de la pièce

Personnage	**Alboury**	**Horn**	**Cal**	**Léone**
Pitch (%)	90	100	110	120

Notons que la voix de Horn est presque identique à celle de l'acteur. Pour cette raison, et comme il s'agit du premier personnage à entrer en scène, à partir d'octobre 2020, c'est la voix naturelle de l'acteur qui est utilisée, avec un simple délai temporel légèrement inférieur à celui des autres voix, de manière à habituer les spectateurs à ce décalage.

6.2.8 Comparaison avec la méthode PSOLA

Afin de décrire les caractéristiques des voix transformées par cet algorithme, nous allons le comparer à une autre méthode de pitch-shifting, parmi les plus connues : la méthode PSOLA (*Pitch Synchronous Overlap-Add*).

6.2.8.1 Présentation de la méthode PSOLA

La méthode PSOLA est illustrée en Figure 125. Elle consiste à analyser le signal pour en repérer les maximums locaux, qui correspondent aux impulsions provoquées par la fermeture de la glotte. Le signal est ensuite découpé en « grains » de parole, centrés sur chaque marqueur de maximum local. Puis, une nouvelle liste de marqueurs est générée, de manière à obtenir le pitch souhaité. À chaque nouveau marqueur est associé le marqueur le plus proche dans le signal initial. Enfin, les grains de parole sont fusionnés, à l'aide d'une OLA.

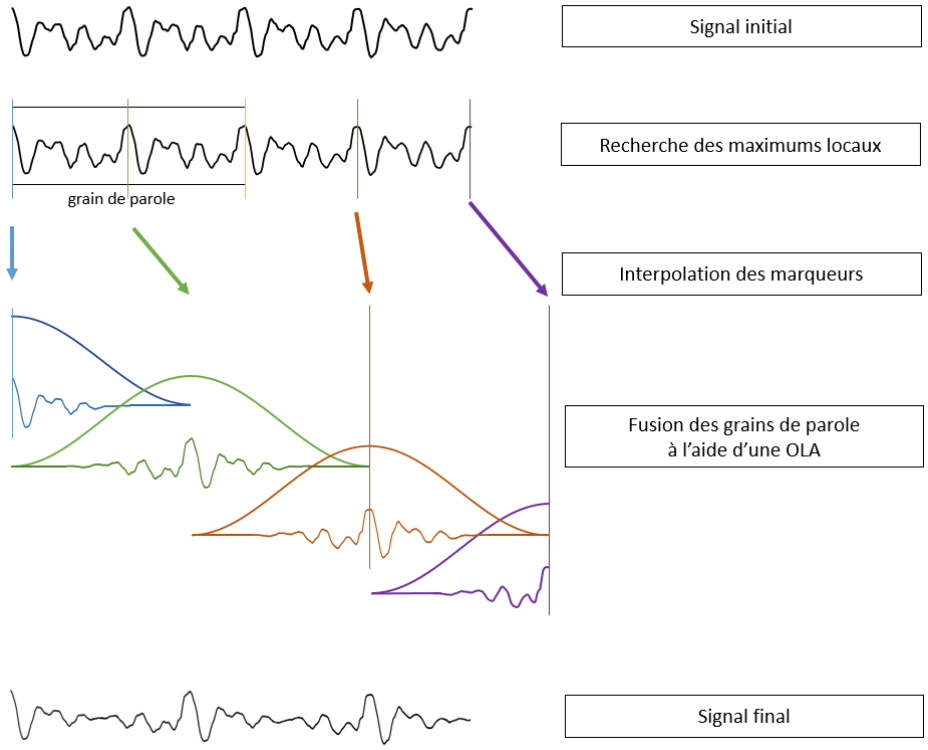


Figure 125 : Méthode PSOLA

6.2.8.2 Conservation du timbre

Les résultats obtenus avec la PSOLA sont assez différents de ceux de notre algorithme, comme le montre l'exemple fourni en Figure 126. Dans les deux cas, le pitch diminue, car l'écart entre chaque impulsion glottique augmente. Cependant, on constate que la forme d'onde varie d'une méthode à l'autre. Ainsi, la forme d'onde est dilatée par le time-stretching ; cela signifie que toutes les fréquences du signal ont diminué, pas seulement la fréquence fondamentale. Au contraire, la PSOLA conserve parfaitement certaines parties de la forme d'onde, mais ajoute également des morceaux de signaux inédits, issus de la fusion des différents grains de parole.

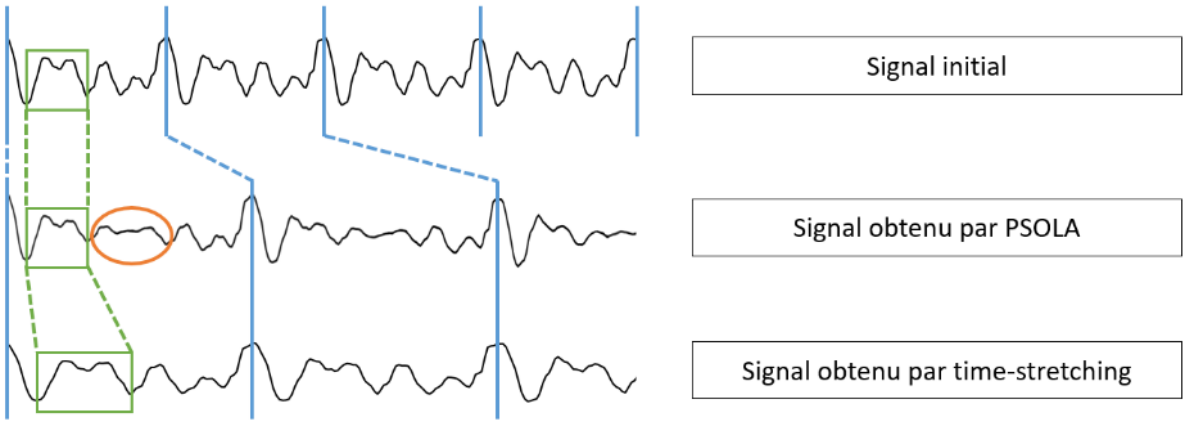


Figure 126 : Comparaison des résultats obtenus avec les deux méthodes présentées

Ainsi, l'intérêt de la PSOLA est de pouvoir modifier le pitch d'une voix, tout en conservant son timbre. Autrement dit, elle permet de simuler une vibration plus rapide ou plus lente des cordes vocales, mais pour un conduit vocal unique. Notre méthode en revanche agrandi / rétrécis à la fois les cordes vocales et le conduit vocal, car elle affecte toutes les fréquences du signal en les multipliant par une valeur donnée. Cette méthode est donc plus efficace pour créer des voix à la fois réalistes (obtenues avec une faible variation de pitch) et reconnaissables (suffisamment différentes de la voix originale).

6.2.8.3 Robustesse

La méthode PSOLA repose sur une détection des impulsions glottiques. Or, celle-ci peut être difficile à réaliser, en particulier lorsque l'enregistrement est bruité. Elle nécessite donc la mise en place d'un détecteur de pitch sophistiqué, dont les paramètres doivent être soigneusement choisis en fonction des signaux traités. Par exemple, il peut être intéressant de savoir à l'avance si la voix modifiée est une voix aiguë (de pitch élevé) ou grave (de pitch faible). Les parties non voisées du signal (pour lesquelles le pitch n'existe pas) ne sont pas traitées.

Au contraire, notre méthode fonctionne de manière systématique : elle traite le signal quelle que soit sa nature voisée ou non voisée. Ses seuls paramètres concernent la taille des fenêtres et les marges de tolérance pour le collage et le rééchantillonnage des fenêtres. Une fois fixés, ces paramètres fourniront de bons résultats, quelle que soit la modification de pitch choisie, ou le type de voix traitée. Elle est donc bien plus robuste que la PSOLA, et permet de modifier des signaux plus variés : des voix bruitées ou chuchotées par exemple, ou même un mélange de plusieurs voix, d'instruments de musique…

6.3 Co-construction d'une sculpture connectée pour télécommander les changements de voix

Lors des premières répétitions en septembre 2016, les changements de voix étaient commandés en régie. Pendant que l'acteur jouait sur scène, je suivais le texte de la pièce, sélectionnant au fur et à mesure la voix de chaque personnage. D'un point de vue technique, le dispositif fonctionnait parfaitement. Cependant, il n'était pas encore intégré à la mise en scène. En particulier, le fait que la voix provienne de deux sources différentes (l'acteur et un haut-parleur) posait problème, car le public pouvait entendre les deux voix, décalées dans le temps d'environ 250 ms. Écouter la voix transformée devait faire l'objet d'un choix conscient de leur part. Par ailleurs, il est apparu nécessaire de concevoir un outil pour permettre à l'acteur de contrôler lui-même ses changements de voix. Cet outil a pris la forme d'une sculpture connectée, permettant à l'acteur de télécommander à distance le logiciel de transformation de voix.

Dans cette partie, nous décrirons les différentes étapes de co-construction de cette sculpture, ainsi que l'évolution des programmes numériques qui lui sont associés.

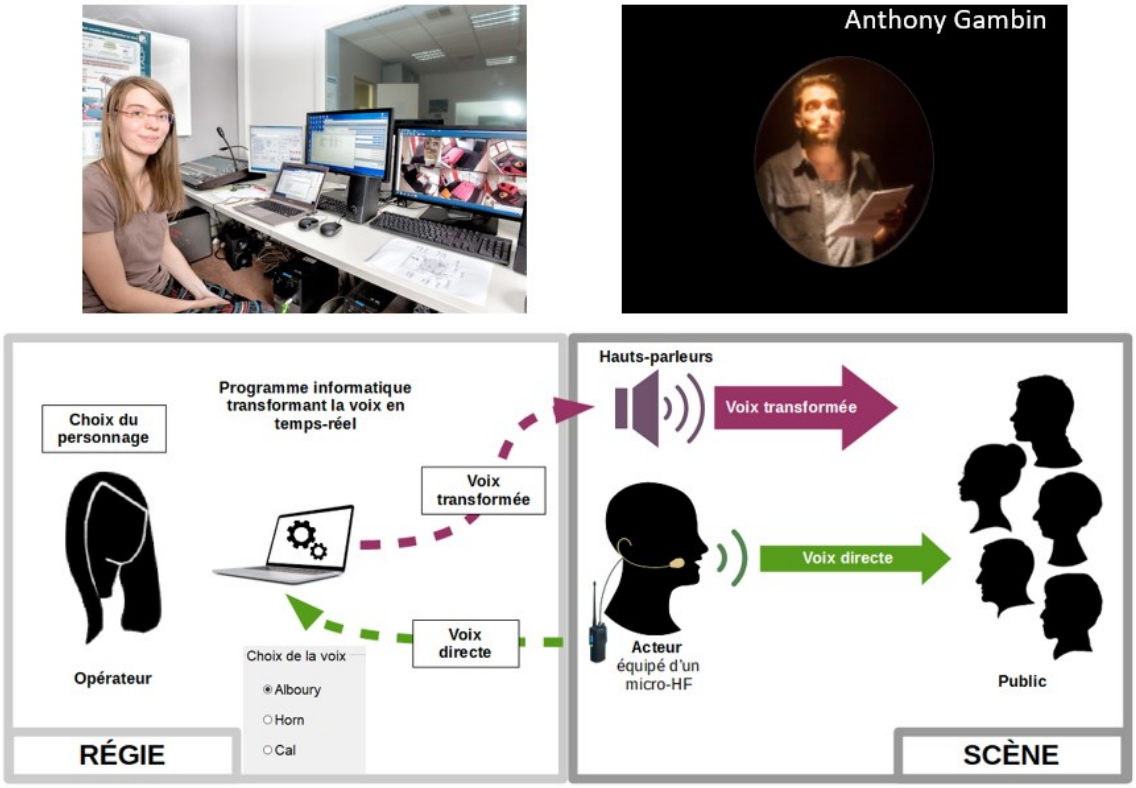


Figure 127 : Dispositif utilisé au cours des premières répétitions (septembre 2016)

6.3.1 Principe de la télécommande

Un premier prototype a été conçu au cours du stage de Natacha Borel (Borel 2017), avec les ressources du fablab universitaire FabMSTIC. Il s'agit d'une télécommande à quatre boutons poussoirs, un pour chaque personnage de la pièce. Son circuit électronique est présenté en Figure 128. Les quatre boutons sont reliés à un microcontrôleur wifi ESP8266, alimenté par

une petite batterie lithium. Lorsque le microcontrôleur est alimenté, il se connecte automatique au réseau wifi, et à l'ordinateur sur lequel est lancé le programme de transformation de voix. Chaque fois qu'on appuie sur un des boutons, l'information est envoyée à l'ordinateur, qui interrompt le processus de transformation de voix, puis le relance avec les paramètres correspondant à la voix sélectionnée.

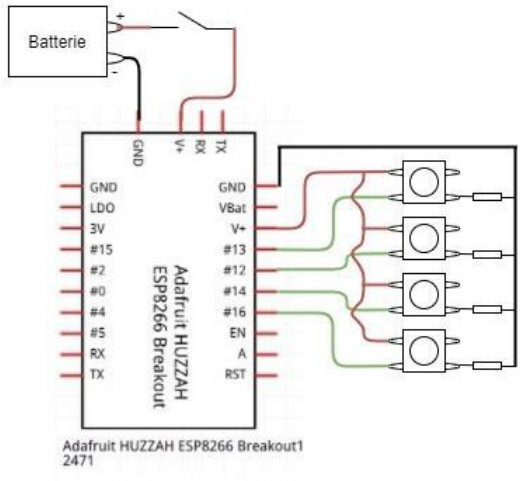


Figure 128 : Schéma électronique du premier prototype de sculpture connectée extrait de (Borel 2017).

6.3.2 Étapes de conception de la sculpture

Alain Quercia a d'abord proposé un dessin du sceptre connecté (cf. Figure 129 a). Le pommeau évoque à la fois un crâne, et un ancien modèle de microphone. Sur le manche, apparaissent les boutons des quatre voix. Il a ensuite réalisé une maquette du crâne en polypropylène (b), que nous avons scanné au laboratoire G-SCOP avec l'aide de Philippe Rene Marin. Le modèle 3D a été corrigé et simplifié par Natacha Borel et Bastien Scher (c), puis imprimé au fabMSTIC (d). Le manche du premier prototype a été conçu directement en 3D par Natacha Borel. Une bande de leds était placée dans le crâne, et changeait de couleur pour chaque personnage. Ce changement de couleur a été abandonné par la suite.

Figure 129 : Évolution de la sculpture connectée
a) ébauche b) maquette en polypropylène c) modèle 3D d) premier prototype e) second prototype f) objet final

Les différentes parties du premier prototype étaient collées ensembles, ce qui rendait difficile l'accès aux composants électroniques. En outre, le manche était très épais, car il contenait le microcontrôleur et la batterie, ce qui n'était ni esthétique, ni ergonomique. Par la suite, nous avons donc choisi de déplacer les composants électroniques dans le crâne pour pouvoir affiner le manche (e). Un nouveau système d'assemblage des différentes parties du sceptre a également été proposé : désormais, les deux moitiés du crâne sont vissées sur le manche. Plusieurs formes de manche ont été testées, suivant les propositions d'Alain (cf. Figure 130). Finalement, le sceptre a évolué en canne : le manche est à présent constitué de deux cylindres, un épais dans lequel est vissé le crâne, et un plus fin, qui peut être vissé sur un bâton sculpté par Alain. Les dimensions du crâne ont alors été repensées, pour correspondre à celle d'un pommeau de canne. Une fois que nous étions satisfaits du prototype, le crâne a été poli par Alain, et un moulage en étain a été réalisé (cf. Figure 129 f).

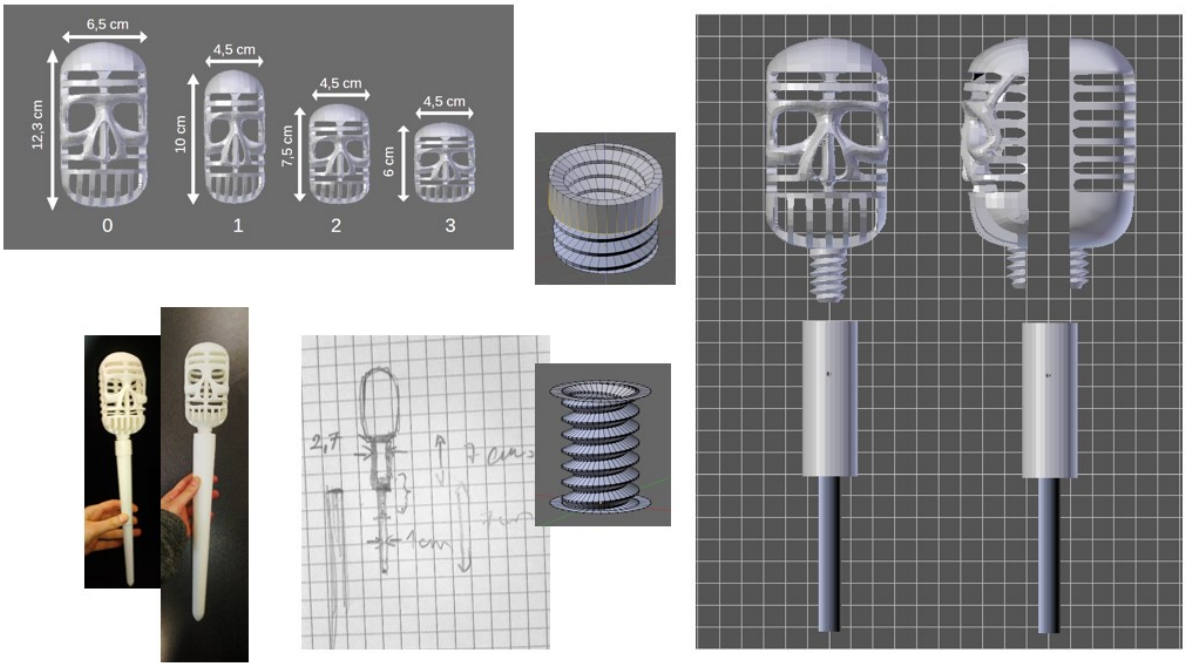


Figure 130 : Différentes propositions pour la forme du sceptre

Les commandes ont également évoluées, car il était difficile pour l'acteur de manipuler les quatre boutons, situés les uns sur les autres. Nous avons proposé de les remplacer par des boutons tactiles, en nous inspirant des capteurs capacitifs pour Arduino décrits par (Badger [sans date]). Chacun de ces boutons consiste en une surface conductrice, reliée à petit circuit électronique très simple. Ils peuvent donc être très facilement déplacés sur le manche, ce qui a permis de tester plusieurs positions. Notre idée initiale était que dans la version finale du sceptre, ces boutons tactiles seraient recouverts par du tissu, et que des repères seraient placés sur le manche pour pouvoir les trouver facilement. Cependant, ces capteurs restaient difficiles à utiliser et les fausses manipulations étaient courantes.

Afin de simplifier le système, nous avons décidé de réduire le nombre de voix disponibles : l'acteur ne peut choisir qu'entre les deux personnages d'une scène. L'interface du logiciel de transformation de voix a donc été modifiée pour permettre à un régisseur de sélectionner le couple de personnages présents à chaque scène. À partir de ce moment, il devenait envisageable d'avoir un sceptre avec un seul bouton, permettant de passer d'une voix à l'autre. Cependant, avec un simple bouton poussoir, l'acteur aurait été incapable de savoir quelle voix était sélectionnée à un moment donné. Nous avons donc fini par opter pour un interrupteur à levier, qui permet de choisir entre deux positions : haute pour la voix aigüe, basse pour la voix grave.

La dernière version de l'interface affiche les informations de connexion du sceptre et permet à un régisseur de changer manuellement la voix sélectionnée en cas de problème avec le sceptre (cf. Figure 131).

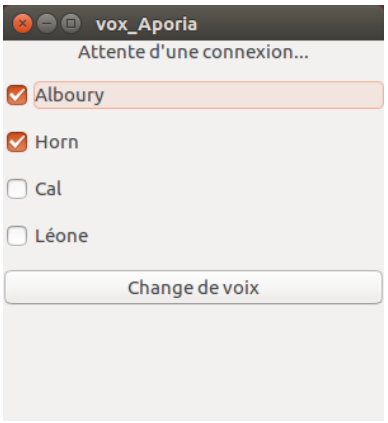


Figure 131 : Interface du logiciel de transformation de voix

6.3.3 Intégrer les transformations de voix à la mise en scène

L'utilisation d'un logiciel de transformation de voix soulève plusieurs problématiques, à la fois pour l'acteur et pour le public présent sur scène. Une partie du développement du sceptre a donc consisté à chercher des solutions pour y répondre au mieux.

6.3.3.1 Latence

Le principal souci provient de la latence entre la voix modifiée et la voix de l'acteur. En effet, bien que le traitement se fasse en temps-réel, il ne peut pas être instantané : quelques dizaines de millisecondes d'enregistrement sont nécessaires pour avoir suffisamment de matière pour réaliser le time-stretching. À cette latence algorithmique incompressible s'ajoute une latence matérielle, liée au transport du signal : il s'agit essentiellement du temps nécessaire pour réaliser la conversion analogique numérique.

Cette latence est particulièrement dommageable. Ainsi, entendre sa voix décalée dans le temps peut provoquer un bégaiement chez l'acteur, tandis que les spectateurs peuvent être gênés par la présence des deux voix sur scène. Même en supposant que la voix modifiée est bien plus

forte que la voix directe, des erreurs de compréhension peuvent apparaître du fait de la désynchronisation entre la voix modifiée et les mouvements des lèvres de l'acteur (cf. effet McGurk). Afin de résoudre ce problème, il a été proposé de faire porter à l'acteur un masque, afin de dissimuler sa bouche et d'atténuer sa voix directe. Si cette solution n'a pas été retenue, plusieurs implémentations de l'algorithme ont été réalisées dans le but de réduire la latence au minimum.

Lors de la première résidence en septembre 2016, ce retard était d'environ 250 ms. L'algorithme était alors implémenté en Matlab. Afin de réduire la latence et de disposer d'un programme libre de droits, une nouvelle implémentation a été réalisée en C, dans un environnement Windows avec l'IDE Codeblocks. La librairie audio OpenAL avait été choisie car elle est multiplateforme, ce qui aurait permis de faire fonctionner l'algorithme aussi bien dans un environnement Windows que Linux ou Mac. Cependant, cette nouvelle implémentation n'a pas permis de réduire les délais : en reprenant des enregistrements effectués en juillet 2017, nous estimons que le délai était toujours entre 250 et 300 ms. Pour pouvoir réduire encore ce délai, il était nécessaire de modifier les paramètres de configuration des cartes sons. L'environnement Linux était donc préférable : à partir de novembre 2017, le logiciel était installé sur une tablette Ubuntu. À partir de ce moment, l'utilisation de la librairie OpenAL n'avait plus d'intérêt : en juillet 2018, une dernière implémentation, basée sur la librairie Pulseaudio, a permis de réduire encore la latence. Elle était d'environ 100 ms lors de la première d'octobre 2018. Après une nouvelle configuration des cartes sons suggérée par James Léonard, elle est désormais de 80 ms.

6.3.3.2 Plusieurs sources sonores

La voix modifiée est nécessairement diffusée via un ou des haut-parleur(s). Elle ne provient donc pas de l'acteur lui-même, ce qui pourrait troubler les spectateurs. Par exemple, on pourrait voir l'acteur devant soi, et entendre la voix de ses personnages venir de la gauche. Certes, il est possible de simuler la direction d'une source sonore uniquement à l'aide de deux haut-parleurs, mais à condition que l'auditeur soit à une position précise. Dans notre cas, l'acteur et les spectateurs sont libres de se déplacer sur la scène, ce qui complique le problème. La solution envisagée à l'heure actuelle consiste à obtenir une diffusion sonore la plus homogène possible sur le plateau, à l'aide d'une grille de plusieurs haut-parleurs. Le système de multidiffusion acoustique RN2i du Gipsa-lab, a ainsi été testé en juillet et décembre 2019, avec l'aide de Jérôme Villeneuve et James Leonard.

6.3.3.3 Effet d'écho

Le microphone porté par l'acteur ne capte pas uniquement sa voix directe, mais également la voix modifiée qui sort des haut-parleurs. Heureusement, la latence empêche tout effet de larsen : le signal ne peut pas être amplifié à l'infini. En revanche, il y a un effet d'écho : les spectateurs peuvent entendre non pas deux, mais trois voix. La troisième voix correspond à l'écho de la voix modifiée, transformée une deuxième fois par le logiciel. Cet écho est suffisamment faible pour ne pas gêner la compréhension lorsque l'acteur parle, mais est perceptible au moment des silences. Il peut être atténué à l'aide d'un *gating* : il s'agit de couper le son qui sort des haut-parleurs dès que son intensité descend en dessous d'un seuil défini à l'avance.

6.4 Conclusion

Nous avons présenté la méthode par laquelle nous avons développé pour le spectacle Aporia un outil numérique, par un processus de co-construction entre un artiste, et une équipe de recherche. Cette collaboration s'est effectuée à travers plusieurs allers-retours entre le fablab, le Living Lab, et les scènes de théâtre. Elle a permis de concevoir un logiciel permettant de transformer des voix de façon réaliste dans l'art de la scène sans altérer leurs propriétés socio-affectives qui, comme nous nous y attendions, se révèlent primordiales : à l'écoute des voix des différents personnages, on identifie toujours la trace de la voix de l'acteur, de la même manière qu'un locuteur garde toujours sa propre voix, lorsqu'il change de rôle social.

Pour poursuivre cette recherche Arts-Sciences sur l'altérité, il est prévu d'intégrer au spectacle un robot de téléprésence : c'est-à-dire une machine téléopérée, perçue comme un autre, ni humain, ni animal. Un premier essai a déjà eu lieu avec le robot Robair, pendant la première d'octobre 2018. Cependant, l'apparente fragilité de la forme actuelle de ce robot et ses déplacements n'étaient pas cohérents avec la tonalité très inquiétante de la pièce. Cela ouvre de nouvelles pistes de recherche, du côté du design et de la navigation sociale, pour étudier ce qui permet de susciter une impression de fragilité ou de robustesse.

CONCLUSION ET PERSPECTIVES

L'objectif d'un robot de téléprésence n'est pas seulement de permettre à une personne de communiquer à distance, mais surtout de conserver un lien social avec ses interlocuteurs, en lui permettant d'être présente dans le même environnement, et de participer à leurs activités. Lorsqu'on développe de tels robots, il est donc essentiel d'étudier le **toucher social**, c'est-à-dire les signaux vecteurs du lien social, et la manière dont ils sont produits et perçus en téléprésence.

Dans cette thèse, nous avons restreint notre étude du toucher social aux signaux vocaux, et en particulier à la **portée vocale**. Pendant une interaction en présentiel, un locuteur adapte en permanence sa portée vocale aux conditions adverses de l'environnement (distance des interlocuteurs et bruit ambiant) pour pouvoir communiquer et échanger ses signaux socio-affectifs. Cependant, la portée vocale peut être radicalement **modifiée** en téléprésence, car la voix doit être enregistrée pour pouvoir ensuite être diffusée dans un nouvel environnement. Cette transmission seule est déjà source d'artefacts (latence, altération du timbre, glitches audio…), sans compter les transformations numériques supplémentaires qui peuvent être ajoutées pour modifier le signal vocal (ex : transformation de pitch). La question du volume sonore est particulièrement importante, puisqu'il est possible d'amplifier une voix de faible portée vocale pour la rendre audible à grande distance ; ou inversement, de diminuer la puissance d'un cri, jusqu'à ce que seul un auditeur situé à proximité du robot puisse l'entendre. Or, une mauvaise portée vocale peut avoir des conséquences sociales importantes : si la voix qui sort du robot est trop forte ou trop faible, elle peut déranger les interlocuteurs, ou rendre l'interaction difficile. Il est donc absolument nécessaire de saisir ce qui constitue une bonne portée vocale, et de comprendre comment cette définition varie en fonction du **contexte** : c'est l'objectif principal de cette thèse.

Après un état des lieux de la robotique de téléprésence actuelle (chapitre 1), nous avons proposé de **définir la portée vocale** comme zone de l'espace délimitée par une limite haute d'intelligibilité, et une limite basse de confort subjectif (chapitre 2). Ces deux limites étant très difficiles à estimer, notre étude de la portée vocale s'est faite de manière indirecte, à travers la perception spatiale acoustique (chapitre 4) ou la mesure des variations de la force de voix (chapitre 5). Pour mener nos expérimentations, nous nous sommes appuyés sur une méthodologie écologique, c'est-à-dire conçue pour reproduire en laboratoire des conditions permettant d'observer des comportements humains les plus proches possibles de la réalité (chapitre 3). Cette méthodologie a également servi au développement d'un outil de transformation de voix pour le spectacle Aporia (chapitre 6).

Nos premières expériences ont porté sur le lien entre **toucher vocal et proxémie** (chapitre 4). Inspirés des études en localisation sonore, nous avons inventé un nouveau protocole, dans le but de quantifier l'effet d'une variation de la distance sociale exprimée par un locuteur sur la perception spatiale acoustique d'un auditeur. Le locuteur était une expérimentatrice entraînée à la production de corpus pour l'étude de la prosodie socio-affective. Elle se déplaçait dans la salle d'expérimentation et prononçait des mots en suivant des consignes concernant sa position par rapport au sujet (direction, distance et orientation) et ses productions vocales (socio-affect et intensité). À chaque nouveau mot, l'auditeur devait indiquer à l'aveugle sa position spatiale. L'expérience a été répétée trois fois : deux fois en présentiel, et une fois à distance, à l'aide d'enregistrements effectués avec notre robot de téléprésence.

Lors de la **première expérience**, nous avons utilisé une mise en scène sophistiquée, de manière à empêcher les sujets de comprendre que la locutrice était une expérimentatrice, et qu'elle se forçait à produire des types de voix très précis : les sujets étaient convaincus qu'ils passaient une expérience sur la perception du goût et de l'odorat, aux côtés d'une experte du domaine. Nous avons alors constaté que la perception des sujets était effectivement influencée par le type de voix utilisé : en effet, les taux de reconnaissance de l'orientation et de la distance variaient en fonction de l'intensité et/ou du socio-affect. Le résultat le plus net concernait la perception de l'orientation : les sujets avaient tendance à percevoir que la locutrice leur tournait le dos, lorsqu'elle utilisait une intensité faible, ou exprimait un socio-affect associé à une petite distance sociale. Nous avons également constaté que le taux de reconnaissance de la distance était plus élevé pour le socio-affect associé à une courte distance sociale. Les variations d'intensité produites par la locutrice n'avaient pas d'effet sur le taux de reconnaissance de la distance. En revanche, la diminution d'intensité induite par l'orientation de dos avait une influence sur la perception de la distance, car les sons produits de dos étaient perçus plus loin en moyenne que les sons produits de face.

Nous avons ensuite **répété l'expérience**, en changeant simplement la manière dont elle était présentée aux sujets : cette fois, ils étaient informés de l'existence des différents types de voix, et devaient même essayer de les reconnaître. L'objectif était de vérifier si les sujets étaient capables dans ces conditions de s'adapter aux variations vocales de la locutrice. Effectivement, les résultats de cette seconde expérience sont significativement différents de ceux de la première : cette fois, le type de voix utilisé n'a aucun effet significatif sur la perception des sujets. Nous avons ainsi confirmé un des principes méthodologiques que nous nous étions fixés : le détournement de l'attention via l'utilisation d'une tâche prétexte, dans le but d'empêcher les sujets de s'observer en train de réaliser la tâche, ce qui modifie leur perception et leur comportement.

Enfin, la **troisième expérience** consistait en un test perceptif en ligne, destinés à comparer les résultats précédents, obtenus en présentiel sur un petit nombre de sujets, à un panel plus large et dans des conditions similaires à celles de la téléprésence. À nouveau, nous avons constaté que le type de voix utilisé influençait la perception des sujets. Évidemment, l'intensité a une importance prédominante, puisqu'une intensité faible est plus souvent associée à un éloignement, une orientation de dos, ou au fait que la locutrice se trouve derrière le robot. L'effet du socio-affect est plus subtil, et difficile à isoler, car il n'est pas indépendant de

l'intensité : en particulier, une voix *breathy* est plus facile à produire et à percevoir lorsqu'elle est associée à une intensité faible. Nous avons constaté que l'effet du socio-affect allait toujours dans le sens de l'intensité « naturelle » : les voix douces sont perçues comme les voix faibles, et les voix distantes comme les voix fortes. Il semble donc que dans la perception de la portée vocale, ce sont bien les qualités acoustiques des signaux qui priment, et non la distance sociale exprimée.

Par ailleurs, nous nous sommes intéressés aux variations de la portée vocale en téléprésence ubiquïte, en étudiant l'impact de l'**effet Lombard** sur l'intensité vocales des locuteurs (chapitre 5). Pour cette nouvelle expérience, nous avons à nouveau eu recours à une mise en scène, en présentant aux sujets une fausse expérience, de sorte qu'ils ne prêtent pas attention aux bruits diffusés et destinés à susciter la parole Lombard.

Cette expérience consistait en un échange de questions / réponses entre deux sujets, l'un face au robot de téléprésence (R), et l'autre en position de pilote (P). La liste de questions était confiée au sujet R ; toutes les 10 questions environ apparaissait une consigne, demandant au sujet P de manipuler l'interface du robot. Nous prétendions que le but de l'expérience était d'étudier la prise en main de l'interface du robot, et que le temps de réponse aux questions nous servait d'étalon. En réalité, différents bruits étaient diffusés pendant l'expérience, soit dans l'environnement du pilote (condition B), soit dans le casque du pilote (condition C) soit dans l'environnement du robot (condition D). L'objectif était de comparer les variations d'intensité des sujets entre la condition A (silence) et les trois conditions de bruit.

Nous avons effectivement observé un effet Lombard : en particulier, les sujets P avaient tendance à parler plus fort lorsqu'ils entendaient du bruit, peu importe que ce bruit soit audible (condition D) ou inaudible par leur interlocuteur (conditions B et C) ; tandis que les sujets R haussaient la voix uniquement en condition D. Ces variations d'intensité étaient significatives, mais faibles, de l'ordre de 1 à 3 dB seulement. Pourtant il semble bien qu'elles soient perceptibles par les sujets, même inconsciemment, puisque nous avons également noté une légère augmentation de l'intensité vocale de sujets P en condition D, c'est-à-dire la condition où les sujets R parlaient également plus fort. Cela suggère l'existence d'un effet d'entraînement combiné à l'effet Lombard.

Enfin, les travaux que nous avons réalisés pour le spectacle Aporia (chapitre 6) nous ont permis de mettre en œuvre une méthode d'innovation technologique en boucles agiles, basées sur une co-construction pluridisciplinaire entre un artiste plasticien, des gens du spectacle, et des chercheurs en robotique. Il s'agissait de concevoir un outil pour aider l'unique acteur de la pièce à incarner différents personnages en modifiant son apparence vocale, tout en conservant son toucher social. Cet outil repose sur un algorithme de modification de pitch que nous avons adapté au temps-réel et implémenté sur une tablette convertible. Ce programme informatique est commandé à distance via une sculpture connectée, développée au fablab fabMSTIC et testée au Living Lab Domus, ainsi que dans plusieurs salles de spectacles.

Passons à présent aux **perspectives** de cette thèse. Tout d'abord, rappelons que nous n'avons trouvé que très peu de travaux dans la littérature qui traitent directement cette notion de portée vocale. En effet, l'étude de la portée vocale se trouve à la frontière de deux domaines assez peu étudié : celui de la force de voix et celui de la proxémie, c'est-à-dire la manière dont les personnes occupent socialement l'espace. Notre thèse peut donc être un bon point de départ pour de futures recherches sur le sujet.

De plus, la question qui nous intéresse, qui est celle du lien entre portée vocale et socio-affect, est très difficile à étudier dans le cadre d'une interaction standard en présentiel, puisque ces deux aspects sont quasiment indissociables l'un de l'autre. Ce n'est que dans le cas où la voix est enregistrée, que sa portée vocale peut être arbitrairement modifiée. Le robot de téléprésence est donc un bon instrument de recherche, car il permet de créer une situation d'interaction réaliste dans laquelle les interlocuteurs ne partagent pas le même environnement, et ne peuvent se percevoir l'un l'autre qu'à travers une interface numérique. Il devient alors possible de modifier leur perception, sans qu'il s'en rendre compte, ce qui ouvre des perspectives intéressantes pour l'étude des boucles sensori-motrices (perception-action).

Par ailleurs, cette thèse nous permet de formuler des recommandations concernant la manière de permettre aux pilotes des robots de téléprésence de maîtriser leur portée vocale. Ce contrôle ne doit en aucun cas être automatisé, car il fait partie du toucher social : c'est au pilote de choisir à chaque instant les limites de sa portée vocale. On peut cependant imaginer des outils pour l'aider à percevoir l'environnement du robot. Ainsi, dans un premier temps, nous avions commencé à réfléchir à une interface visuelle, qui représenterait en vue du dessus la portée vocale du robot. Pour la mettre en œuvre, il était cependant nécessaire de disposer d'une « boîte-noire », capable à chaque instant de calculer la portée vocale adaptée au contexte acoustique et social, à partir des signaux enregistrés par les capteurs robot. Ce problème n'est pas facile à résoudre dans le cas général : une modélisation acoustique peut rapidement devenir très complexe si l'on doit tenir compte des réverbérations dans l'environnement ; de plus, cela suppose d'être capable de prendre en compte un ensemble mouvant de règles sociales, susceptible d'évoluer de lieu en lieu, voire d'heure en heure. Dans le cadre scolaire par exemple, ces règles ne sont pas les mêmes dans la cours de récréation que dans les salles de classe, aux heures de pause et aux heures de cours. Pour être véritablement utile, une telle interface devrait être réalisée à travers une démarche de type *Living Lab*, à laquelle participeraient les utilisateurs réguliers de ces robots. En absence de tels utilisateurs, nous avons préféré nous concentrer sur l'expérimentation, plutôt que sur le développement de cette interface.

Avec le recul, nous pouvons malgré tout fixer quelques objectifs concernant le toucher vocal en robotique de téléprésence. Tout d'abord, la perception que le pilote a de l'environnement du robot doit être la plus proche possible de celle qu'il aurait s'il était présent sur place. En particulier, il doit pouvoir repérer dans l'espace la position des sources sonores, au moins approximativement, ne serait-ce que pour pouvoir compenser les limites de son champ de vision. Ceci peut être réalisé très simplement, à l'aide d'une écoute binaurale rudimentaire qui consiste à utiliser un couple de micro, placé de part et d'autre du robot. De plus, la voix de la personne qui pilote le robot doit être reproduite de façon fidèle : en particulier, il faut s'assurer qu'elle soit aussi forte dans l'environnement local que dans l'environnement distant. Cela n'est

pas toujours possible avec un microphone embarqué dans l'ordinateur du pilote, car le rapport signal sur bruit peut être trop faible : mieux vaut donc utiliser un micro-casque, pour capter la voix au niveau de la bouche.

Pour finir, notons que l'épidémie de Covid19 a fait émerger brutalement la notion de distanciation sociale, et s'est accompagnée d'une transition forcée au « tout-numérique » partout où le télétravail était possible, y compris dans les domaines de la santé et de l'éducation. En conséquence, le sujet du toucher social en téléprésence paraît soudain moins exotique. Nous espérons qu'il connaîtra un développement important dans les années à venir, et que de nouveaux travaux se pencheront plus précisément sur les aspects moins abordés dans cette thèse, que sont le toucher visuel et tactile. La robotique de téléprésence pourrait alors permettre de lutter efficacement contre l'isolement social. Bien utilisée, elle aurait également un intérêt écologique, en limitant les transports longue distance ; à condition que ces robots soient durables et faciles à réparer.

BIBLIOGRAPHIE

ABAD, C., FEARDAY, A. et SAFDAR, N., 2010. Adverse effects of isolation in hospitalised patients: a systematic review. In : *The Journal of Hospital Infection*. octobre 2010. Vol. 76, n° 2, pp. 97-102. DOI 10.1016/j.jhin.2010.04.027.

AINSWORTH, Mary D Salter et BOWLBY, John, 1991. An Ethological Approach to Personality Development. In : *American psychologist*. avril 1991. Vol. 46, n° 4, pp. 333-341.

AMAZI DEBORAH K. et GARBER SHARON R., 1982. The Lombard Sign as a Function of Age and Task. In : *Journal of Speech, Language, and Hearing Research*. 1 décembre 1982. Vol. 25, n° 4, pp. 581-585. DOI 10.1044/jshr.2504.581.

ANON., 2003a. *UIT-T G.114 : Temps de transmission dans un sens* [en ligne]. S.l. Union internationale des télécommunications. [Consulté le 4 janvier 2019]. Disponible à l'adresse : http://handle.itu.int/11.1002/1000/6254.

ANON., 2003b. *UIT-T G.131 : Echo pour le locuteur et sa réduction* [en ligne]. S.l. Union internationale des télécommunications. [Consulté le 4 janvier 2019]. Disponible à l'adresse : https://www.itu.int/rec/T-REC-G.131/fr.

ANON., 2015. *UIT-T G.107 : The E-model : a computational model for use in transmission planning* [en ligne]. S.l. Union internationale des télécommunications. [Consulté le 4 janvier 2019]. Disponible à l'adresse : https://www.itu.int/rec/T-REC-G.107-201506-I/fr.

ANON., 2019. Apple peut donner l'illusion que vous regardez votre interlocuteur dans les yeux pendant un FaceTime. In : *Le Monde avec AFP* [en ligne]. 4 juillet 2019. [Consulté le 22 juillet 2019]. Disponible à l'adresse : https://www.lemonde.fr/pixels/article/2019/07/04/facetime-apple-peut-donner-l-illusion-que-vous-regardez-votre-interlocuteur-dans-les-yeux_5485352_4408996.html.

ANON., 2020. *Baromètre des connexions Internet mobiles en France métropolitaine - Année 2019* [en ligne]. S.l. nPerf. [Consulté le 10 mars 2020]. Disponible à l'adresse : https://media.nperf.com/files/publications/FR/2019-02-03_Barometre-connexions-mobiles-metropole-nPerf-2019.pdf.

ARONS, Barry, 1992. A Review of The Cocktail Party Effect. In : *Journal of the American Voice I/O Society*. 1992. Vol. 12, n° 7, pp. 35-50.

AUBERGÉ, Véronique, 1991. *La synthèse de la parole : des règles aux lexiques* [en ligne]. These de doctorat. S.l. : Grenoble 2. [Consulté le 17 juin 2020]. Disponible à l'adresse : http://www.theses.fr/1991GRE29041.

AUBERGÉ, Véronique, 2015. Gestual-facial-vocal prosody as the main tool of the socio-affective "glue": interaction is a dynamic system. In : *International workshop on audio-visual affective prosody in social interaction*. Bordeaux. 2015.

AUDIBERT, Nicolas, 2008. *Prosodie de la parole expressive : dimensionnalité d'énoncés méthodologiquement contrôlés authentiques et actés* [en ligne]. phdthesis. S.l. : Institut National Polytechnique de Grenoble - INPG. [Consulté le 4 janvier 2019]. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-00489924/document.

AUGOYARD, Jean-François, AMPHOUX, Pascal et BALAŸ, Olivier, 1985. 07 : *Environnement sonore et communication interpersonnelle Tome 1* [en ligne]. Research Report. S.l. CRESSON, CNT. [Consulté le 18 janvier 2019]. Disponible à l'adresse : https://hal.archives-ouvertes.fr/hal-01373804.

BADGER, Paul, [sans date]. Arduino Playground - CapacitiveSensor. In : [en ligne]. [Consulté le 18 février 2020]. Disponible à l'adresse : https://playground.arduino.cc/Main/CapacitiveSensor/.

BERNARDONI, Nathalie Henrich, 2012. Physiologie de la voix chantée: vibrations laryngées et adaptations phono-résonantielles. In : *40èmes Entretiens de Médecine physique et de réadaptation* [en ligne]. Montpellier, France : s.n. mars 2012. pp. pp.17-32. Disponible à l'adresse : hal-00680692.

BERTHOZ, Alain, 2003. *Physiologie de la perception et de l'action* [en ligne]. Paris. Collège de France. [Consulté le 20 juillet 2020]. Disponible à l'adresse : https://www.college-de-france.fr/media/alain-berthoz/UPL17178_UPL52025_BerthozR01_02.pdf.

BEST, Virginia, CARLILE, Simon, JIN, Craig et VAN SCHAIK, André, 2005. The role of high frequencies in speech localization. In : *The Journal of the Acoustical Society of America*. juillet 2005. Vol. 118, n° 1, pp. 353-363. DOI 10.1121/1.1926107.

BLAUERT, Jens, 1997. *Spatial Hearing: The Psychophysics of Human Sound Localization*. S.l. : MIT Press. ISBN 978-0-262-02413-6.

BLESSER, Barry, 2007. *Presentation to Belmont library* [en ligne]. Belmont library, Belmont, Massachusetts : 2007. [Consulté le 13 février 2019]. Disponible à l'adresse : http://www.blesser.net/downloads/Lecture_edited_64.mp3.

BOREL, Natacha, 2017. *Projet Aporia*. Rapport de stage M2. S.l. DCISS - UGA.

BORIL, Hynek et POLLÁK, Petr, 2005. Design and Collection of Czech Lombard Speech Database. In : *Ninth European Conference on Speech Communication and Technology*. S.l. : s.n. 2005. pp. 1577-1580.

BOTTALICO, Pasquale, 2018. Lombard effect, ambient noise, and willingness to spend time and money in a restaurant. In : *The Journal of the Acoustical Society of America*. 1 septembre 2018. Vol. 144, n° 3, pp. EL209-EL214. DOI 10.1121/1.5055018.

BRUNGART, Douglas S. et SCOTT, Kimberly R., 2001. The effects of production and presentation level on the auditory distance perception of speech. In : *The Journal of the Acoustical Society of America*. juillet 2001. Vol. 110, n° 1, pp. 425-440. DOI 10.1121/1.1379730.

BUXTON, William, 1992. Telepresence: Integrating shared task and person spaces. In : *Proceedings of graphics interface*. S.l. : Canadian Information Processing Society Toronto, Canada. 1992. pp. 123–129.

CABRERA, Densil, SATO, Hayato, MARTENS, William L et LEE, Doheon, 2009. Binaural Measurement and Simulation of the Room Acoustical Response from a Person's Mouth to their Ears. In : *Acoustics Australia*. 2009. Vol. 37, n° 3.

CALCAGNO, Esteban R, ABREGÚ, Ezequiel L, EGUÍA, Manuel C et VERGARA, Ramiro, 2012. The Role of Vision in Auditory Distance Perception. In : *Perception*. février 2012. Vol. 41, n° 2, pp. 175-192. DOI 10.1068/p7153.

CARLILE, Simon, 1996. The Physical and Psychophysical Basis of Sound Localization. In : CARLILE, Simon, *Virtual Auditory Space: Generation and Applications* [en ligne]. Berlin, Heidelberg : Springer Berlin Heidelberg. pp. 27-78. [Consulté le 3 janvier 2019]. ISBN 978-3-662-22596-7. Disponible à l'adresse : http://link.springer.com/10.1007/978-3-662-22594-3_2.

CARLSON, V R, 1977. Instructions and perceptual constancy judgment. In : *Stability and constancy in visual perception: Mecanisms and Processes*. W Epstein. New York : s.n. pp. 217-254.

CESTA, Amedeo, CORTELLESSA, Gabriella, ORLANDINI, Andrea et TIBERIO, Lorenza, 2016. Long-Term Evaluation of a Telepresence Robot for the Elderly: Methodology and Ecological Case Study. In : *International Journal of Social Robotics*. 1 juin 2016. Vol. 8, n° 3, pp. 421-441. DOI 10.1007/s12369-016-0337-z.

CHÉRET, Jules, 1896. *Le théâtrophone* [en ligne]. 1896. S.l. : Imprimerie Chaix, Paris. [Consulté le 26 juillet 2019]. Disponible à l'adresse : https://fr.wikipedia.org/w/index.php?title=Th%C3%A9%C3%A2trophone&oldid=157284988.

CHERMAYEFF, Maro et LE GOFF, Christine, 2017. Soundbreaking, la grande aventure de la musique enregistrée. In : . 2017.

CHU, W. T. et WARNOCK, A. C. C., 2002. Detailed Directivity of Sound Fields Around Human Talkers. In : *National Research Council Canada* [en ligne]. 1 décembre 2002. [Consulté le 4 janvier 2019]. DOI 10.4224/20378930. Disponible à l'adresse : http://nparc.cisti-icist.nrc-cnrc.gc.ca/eng/view/object/?id=c1449aba-ee5b-48de-8312-ec325d31ef37.

COLEMAN, Paul D., 1962. Failure to Localize the Source Distance of an Unfamiliar Sound. In : *The Journal of the Acoustical Society of America*. 1 mars 1962. Vol. 34, n° 3, pp. 345-346. DOI 10.1121/1.1928121.

COUREAU-FALQUERHO, Edwige, SIMONIAN, Stéphane et PEROTIN, Catherine, 2017. Expérimentation « Robot lycéen » en Auvergne Rhône-Alpes. In : . 2017. pp. 8.

DÉMONET, Jean-François, [sans date]. *Cerveau et langage oral* [en ligne]. S.l. : s.n. [Consulté le 19 juillet 2020]. Disponible à l'adresse : https://www.universalis.fr/encyclopedie/cerveau-et-langage-oral/.

DESHAYS, Daniel, 2006. *Pour une écriture du son* [en ligne]. S.l. : s.n. [Consulté le 18 mars 2020]. ISBN 978-2-252-03565-8. Disponible à l'adresse : https://www.klincksieck.com/livre/817-pour-une-ecriture-du-son.

DUPOUX, Emmanuel, 2018. Cognitive Science in the era of Artificial Intelligence: A roadmap for reverse-engineering the infant language-learner. In : *Cognition*. avril 2018. Vol. 173, pp. 43-59. DOI 10.1016/j.cognition.2017.11.008.

EDLUND, Jens, HELDNER, Mattias et GUSTAFSON, Joakim, 2012. Who am I speaking at? Perceiving the head orientation of speakers from acoustic cues alone. In : *LREC*. S.l. : s.n. 2012. pp. 4.

ENZNER, Gerald, BUCHNER, Herbert, FAVROT, Alexis et KUECH, Fabian, 2014. Acoustic Echo Control. In : *Academic Press Library in Signal Processing* [en ligne]. S.l. : Elsevier. pp. 807-877. [Consulté le 16 janvier 2019]. ISBN 978-0-12-396501-1. Disponible à l'adresse : https://linkinghub.elsevier.com/retrieve/pii/B9780123965011000303.

ERIKSSON, Anders et TRAUNMÜLLER, Hartmut, 2002. Perception of vocal effort and distance from the speaker on the basis of vowel utterances. In : *Perception & Psychophysics*. janvier 2002. Vol. 64, n° 1, pp. 131-139. DOI 10.3758/BF03194562.

FELS, Janina et VORLAENDER, Michael, 2008. The next generation of artificial heads. In : *The Journal of the Acoustical Society of America*. 1 mai 2008. Vol. 123, n° 5, pp. 3159-3159. DOI 10.1121/1.2933197.

FELS, Janina et VORLÄNDER, Michael, 2009. Anthropometric Parameters Influencing Head-Related Transfer Functions. In : [en ligne]. avril 2009. [Consulté le 4 février 2020]. Disponible à l'adresse : https://www.ingentaconnect.com/content/dav/aaua/2009/00000095/00000002/art00014.

FITCH, W. Tecumseh, 2018. The Biology and Evolution of Speech: A Comparative Analysis. In : *Annual Review of Linguistics*. 2018. Vol. 4, n° 1, pp. 255-279. DOI 10.1146/annurev-linguistics-011817-045748.

FÓNAGY, Iván, 1983. *La vive voix : essais de psycho-phonétique*. Payot. S.l. : s.n.

FONTAN, Lionel, 2012. *De la mesure de l'intelligibilité à l'évaluation de la compréhension de la parole pathologique en situation de communication* [en ligne]. Linguistique. S.l. : Université Toulouse le Mirail-Toulouse II. Disponible à l'adresse : <tel-00797883>. <NNT : 2012TOU20113>.

FUKUI, Kotaro, KUSANO, Toshihiro, MUKAEDA, Yoshikazu, SUZUKI, Yuto, TAKANISHI, Atsuo et HONDA, Masaaki, 2010. Speech Robot Mimicking Human Articulatory Motion. In : *Interspeech 2010*. S.l. : s.n. 2010. pp. 1021-1024.

FUX, Thibaut, 2012. *Vers un système indiquant la distance d'un locuteur par transformation de sa voix* [en ligne]. thesis. S.l. : Grenoble. [Consulté le 4 janvier 2019]. Disponible à l'adresse : http://www.theses.fr/2012GRENT120.

FUX, Thibaut et ZIMPFER, Véronique, 2009. ISL-RV 218 : *Spatialisation du son : corrélation entre la distance perçue et l'effort vocal*. France. French-German Research Institute of Saint-Louis.

GABRIEL, Claude, 2018. Production de la parole et voix humaine. In : *Cours d'acoustique 2018-2019* [en ligne]. Haute Ecole Libre de Bruxelles : s.n. Cinématographie. [Consulté le 16 juillet 2019]. Disponible à l'adresse : http://www.claudegabriel.be/Acoustique%20chapitre%209.pdf.

GARDNER, Mark B., 1969. Distance Estimation of 0° or Apparent 0°-Oriented Speech Signals in Anechoic Space. In : *The Journal of the Acoustical Society of America*. janvier 1969. Vol. 45, n° 1, pp. 47-53. DOI 10.1121/1.1911372.

GARNIER, Maëva, 2007. *Communiquer en environnement bruyant : de l'adaptation jusqu'au forçage vocal*. Paris VI : Université Pierre et Marie Curie.

GARNIER, Maëva, HENRICH, Nathalie et DUBOIS, Danièle, 2010. Influence of Sound Immersion and Communicative Interaction on the Lombard Effect. In : *Journal of Speech, Language, and Hearing Research*. juin 2010. Vol. 53, n° 3, pp. 588-608. DOI 10.1044/1092-4388(2009/08-0138).

GARNIER, Maëva, MÉNARD, Lucie et ALEXANDRE, Boris, 2018. Hyper-articulation in Lombard speech: An active communicative strategy to enhance visible speech cues? In : *The Journal of the Acoustical Society of America*. août 2018. Vol. 144, n° 2, pp. 1059-1074. DOI 10.1121/1.5051321.

GASIGLIA-LASTER, Danièle, 1983. Splendeurs et misères du théâtrophone. In : *Romantisme*. 1983. Vol. 13, n° 41, pp. 74-78. DOI 10.3406/roman.1983.4655.

GOBL, C, 2003. The role of voice quality in communicating emotion, mood and attitude. In : *Speech Communication*. avril 2003. Vol. 40, n° 1-2, pp. 189-212. DOI 10.1016/S0167-6393(02)00082-1.

GORDON, Michael S. et ROSENBLUM, Lawrence D., 2004. Perception of Sound-Obstructing Surfaces Using Body-Scaled Judgments. In : *Ecological Psychology*. avril 2004. Vol. 16, n° 2, pp. 87-113. DOI 10.1207/s15326969eco1602_1.

GRAMMING, Patricia, SUNDBERG, Johan, TERNSTRÖM, Sten, LEANDERSON, Rolf et PERKINS, William H., 1988. Relationship between changes in voice pitch and loudness. In : *Journal of Voice*. janvier 1988. Vol. 2, n° 2, pp. 118-126. DOI 10.1016/S0892-1997(88)80067-5.

GROTHE, Benedikt, PECKA, Michael et MCALPINE, David, 2010. Mechanisms of Sound Localization in Mammals | Physiological Reviews. In : *Physiological reviews*. 1 juillet 2010. Vol. 90, n° 3, pp. 983-1012.

HAGEN, P., LYONS, G. D. et NUSS, D. W., 1996. Dysphonia in the elderly: diagnosis and management of age-related voice changes. In : *Southern medical journal*. février 1996. Vol. 89, n° 2, pp. 204-207. DOI 10.1097/00007611-199602000-00009.

HALKOSAARI, Teemu et VAALGAMAA, Markus, 2005. Directivity of Artificial and Human Speech. In : *J. Audio Eng. Soc.* 2005. Vol. 53, n° 7, pp. 12.

HALL, Edward T., BIRDWHISTELL, Ray L., BOCK, Bernhard, BOHANNAN, Paul, DIEBOLD, A. Richard, DURBIN, Marshall, EDMONSON, Munro S., FISCHER, J. L., HYMES, Dell, KIMBALL, Solon T., BARRE, Weston La, LYNCH, Frank, J., S., MCCLELLAN, J. E., MARSHALL, Donald S., MILNER, G. B., SARLES, Harvey B., TRAGER, George L et VAYDA, Andrew P., 1968. Proxemics [and Comments and Replies]. In : *Current Anthropology*. 1968. Vol. 9, n° 2/3, pp. 83-108.

HALL, Edward Twitchell, 1966. The hidden dimension. Vol. 609. Garden City, NY: Doubleday.

HARRISON, C.S. et MAIR, G.M., 2007. A mechatronic approach to supernormal auditory localisation for telepresence. In : *Mechatronics*. novembre 2007. Vol. 17, n° 9, pp. 501-510. DOI 10.1016/j.mechatronics.2007.05.005.

HAUSTEIN, B. G., 1969. Hypothesen über die einohrige entfernungswahrnehmung des menschlichen gehörs (hypotheses about the perception of distance in human hearing with one ear). In : *Hochfrequenztech. u. Elektroakustik*. 1969. Vol. 78, pp. 46-57.

HAWLEY, Monica L., SHERLOCK, LaGuinn P. et FORMBY, Craig, 2017. Intra- and Intersubject Variability in Audiometric Measures and Loudness Judgments in Older Listeners with Normal Hearing. In : *OtoRhinoLaryngology by Sfakianakis G.Alexandros* [en ligne]. 10 mars 2017. [Consulté le 29 juillet 2019]. Disponible à l'adresse : https://otorhinolaryngologyblog.wordpress.com/2017/03/10/intra-and-intersubject-variability-in-audiometric-measures-and-loudness-judgments-in-older-listeners-with-normal-hearing/.

HAYAMIZU, Akira, IMAI, Michita, NAKAMURA, Keisuke et NAKADAI, Kazuhiro, 2014. Volume adaptation and visualization by modeling the volume level in noisy environments for telepresence system. In : *Proceedings of the second international conference on Human-agent interaction - HAI '14* [en ligne]. Tsukuba, Japan : ACM Press. 2014. pp. 67-74. [Consulté le 4 janvier 2019]. ISBN 978-1-4503-3035-0. Disponible à l'adresse : http://dl.acm.org/citation.cfm?doid=2658861.2658875.

HCSP, 2013. *Niveaux acceptables d'expositions aux niveaux sonores élevés de la musique* [en ligne]. Paris. Haut Conseil de la Santé Publique. [Consulté le 12 juillet 2019]. Disponible à l'adresse : https://www.hcsp.fr/Explore.cgi/avisrapportsdomaine?clefr=378.

HIRAHARA, Tatsuya, SAWADA, Yuki et MORIKAWA, Daisuke, 2015. Sound Localization of Dynamic Binaural Signals Provided Using a Pinna-Less Dummy Head or a Stereo Microphone. In : *Interdisciplinary Information Sciences*. 2015. Vol. 21, n° 2, pp. 159-166. DOI 10.4036/iis.2015.A.07.

IRCAM, 2002. HRTF Database. In : [en ligne]. 2002. [Consulté le 18 mars 2020]. Disponible à l'adresse : http://recherche.ircam.fr/equipes/salles/listen/index.html.

JANSEN, Bart, GOODWIN, Timothy, GUPTA, Varun, KUIPERS, Fernando et ZUSSMAN, Gil, 2018. Performance Evaluation of WebRTC-based Video Conferencing. In : *SIGMETRICS Perform. Eval. Rev.* mars 2018. Vol. 45, n° 3, pp. 56–68. DOI 10.1145/3199524.3199534.

JOHNSTONE, Tom et SCHERER, Klaus R, 1999. The effects of emotions on voice quality. In : *Proceedings of the XIVth international congress of phonetic sciences*. S.l. : s.n. 1999. pp. 2029--2032.

JUNQUA, J.-C., FINCKE, S. et FIELD, K., 1999. The Lombard effect: a reflex to better communicate with others in noise. In : *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)* [en ligne]. Phoenix, AZ, USA : IEEE. 1999. pp. 2083-2086 vol.4. [Consulté le 3 avril 2019]. ISBN 978-0-7803-5041-0. Disponible à l'adresse : http://ieeexplore.ieee.org/document/758343/.

JUNQUA, Jean-Claude, 1996. The influence of acoustics on speech production: A noise-induced stress phenomenon known as the Lombard reflex. In : *Speech Communication*. 1 novembre 1996. Vol. 20, n° 1, pp. 13-22. DOI 10.1016/S0167-6393(96)00041-6.

KATO, Hiroaki, TAKEMOTO, Hironori, NISHIMURA, Ryouichi et MOKHTARI, Parham, 2010. On the human ability to auditorily perceive human speaker's facing angle. In : *2010 4th International Universal Communication Symposium*. S.l. : s.n. octobre 2010. pp. 387-391.

KIMURA, Atsunobu, IHARA, Masayuki, KOBAYASHI, Minoru, MANABE, Yoshitsugu et CHIHARA, Kunihiro, 2007. Visual Feedback: Its Effect on Teleconferencing. In : JACKO, Julie A. (éd.), *Human-Computer Interaction. HCI Applications and Services* [en ligne]. Berlin, Heidelberg : Springer Berlin Heidelberg. pp. 591-600. [Consulté le 4 janvier 2019]. ISBN 978-3-540-73109-2. Disponible à l'adresse : http://link.springer.com/10.1007/978-3-540-73111-5_67.

KIMURA, Atsunobu, SHIMADA, Yoshihiro et KOBAYASHI, Minoru, 2006. Ambient Pre-Communication. In : CAI, Yang et ABASCAL, Julio (éd.), *Ambient Intelligence in Everyday Life* [en ligne]. Berlin, Heidelberg : Springer Berlin Heidelberg. pp. 142-156. [Consulté le 4 janvier 2019]. ISBN 978-3-540-37785-6. Disponible à l'adresse : http://link.springer.com/10.1007/11825890_7.

KOLARIK, Andrew J., MOORE, Brian C. J., ZAHORIK, Pavel, CIRSTEA, Silvia et PARDHAN, Shahina, 2016. Auditory distance perception in humans: a review of cues, development, neuronal bases, and effects of sensory loss. In : *Attention, Perception, & Psychophysics*. 1 février 2016. Vol. 78, n° 2, pp. 373-395. DOI 10.3758/s13414-015-1015-1.

KORN, T. S., 1954. Effect of Psychological Feedback on Conversational Noise Reduction in Rooms. In : *The Journal of the Acoustical Society of America*. 1954. Vol. 26, n° 5, pp. 793. DOI 10.1121/1.1907420. world

KRISTOFFERSSON, Annica, CORADESCHI, Silvia et LOUTFI, Amy, 2013. A Review of Mobile Robotic Telepresence. In : *Adv. in Hum.-Comp. Int.* janvier 2013. Vol. 2013, pp. 3:3–3:3. DOI 10.1155/2013/902316.

LABOV, William, 1970. *The study of language in its social context.* [en ligne]. Cambridge University Press. University of Pennsylvania : s.n. Disponible à l'adresse : https://doi.org/10.1017/CBO9780511618208.004.

LAMBOURG, Christophe, 2002. Évaluation de l'intelligibilité de la parole dans les établissements recevant du public sonorisés. In : *Acoustique & Techniques*. 2002. Vol. 29.

LANGENDIJK, Erno H. A. et BRONKHORST, Adelbert W., 2000. Fidelity of three-dimensional-sound reproduction using a virtual auditory display. In : *The Journal of the Acoustical Society of America*. janvier 2000. Vol. 107, n° 1, pp. 528-537. DOI 10.1121/1.428321.

LAVER, J, 1980. The Phonetic Description of Voice Quality. In : *The Phonetic Description of Voice Quality*. 1980. Vol. 31, pp. 1-186.

LE GOFF, Emeline et GIORGIS, Zoé, 2019. *L'escape game comme outil d'expérience : un essai dans le domaine de la robotique sociale*. S.l. UFR LLASIC, Département I3L.

LEE, Kwan Min, 2004. Presence, explicated. In : *Communication theory*. 2004. Vol. 14, n° 1, pp. 27–50.

LEEB, Robert, TONIN, Luca, ROHM, Martin, DESIDERI, Lorenzo, CARLSON, Tom et MILLAN, Jose del R., 2015. Towards Independence: A BCI Telepresence Robot for People With Severe Motor Disabilities. In : *Proceedings of the IEEE*. juin 2015. Vol. 103, n° 6, pp. 969-982. DOI 10.1109/JPROC.2015.2419736.

LENAY, Charles, 2010. « C'est très touchant » La valeur émotionnelle du contact. In : *Intellectica*. janvier 2010. Vol. 53, n° 1, pp. 359-397.

LENAY, Charles, 2016. *Le Corps Infini* [en ligne]. ENS Louis-Lumière : 2016. [Consulté le 22 juillet 2019]. Disponible à l'adresse : https://www.youtube.com/watch?v=kcy8B-mH6Bo.

LÉON, Pierre, 1993. *Précis de phonostylistique : parole et expressivité*. Nathan. Paris : s.n. Fac Linguistique.

LIÉNARD, Jean-Sylvain, 1977. *Les Processus de la communication parlée : introduction à l'analyse et à la synthèse de la parole*. Paris; New York : Masson. ISBN 978-2-225-48037-9.

LIÉNARD, Jean-Sylvain, 2018. Représentation et Estimation de la Force de Voix à partir du Spectre Moyen à Long Terme. In : *XXXe Journées d'Etudes sur la Parole* [en ligne]. Aix en Provence : International Speech Communication Association. juin 2018. [Consulté le 16 janvier 2019]. Disponible à l'adresse : https://hal.archives-ouvertes.fr/hal-01871854.

LINDBLOM, B., 1990. Explaining Phonetic Variation: A Sketch of the H&H Theory. In : HARDCASTLE, William J. et MARCHAL, Alain (éd.), *Speech Production and Speech Modelling* [en ligne]. Dordrecht : Springer Netherlands. NATO ASI Series. pp. 403-439. [Consulté le 30 janvier 2019]. ISBN 978-94-009-2037-8. Disponible à l'adresse : https://doi.org/10.1007/978-94-009-2037-8_16.

LIU, Chaoran, ISHI, Carlos et ISHIGURO, Hiroshi, 2019. Auditory scene reproduction for tele-operated robot systems. In : *Advanced Robotics*. 18 avril 2019. Vol. 33, n° 7-8, pp. 415-423. DOI 10.1080/01691864.2019.1599729.

LIU, Chaoran, ISHI, Carlos T. et ISHIGURO, Hiroshi, 2015. Bringing the Scene Back to the Tele-operator: Auditory Scene Manipulation for Tele-presence Systems. In : *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction - HRI '15*

[en ligne]. Portland, Oregon, USA : ACM Press. 2015. pp. 279-286. [Consulté le 4 janvier 2019]. ISBN 978-1-4503-2883-8. Disponible à l'adresse : http://dl.acm.org/citation.cfm?doid=2696454.2696494.

LOMBARD, É., 1911. Le signe de l'élévation de la voix [The sign of voice raising]. In : *Annales des Maladies de l'Oreille et du Larynx*. 1911. Vol. XXXVII, pp. 101-109.

LOOMIS, Jack M., KLATZKY, Roberta L., PHILBECK, John W. et GOLLEDGE, Reginald G., 1998. Assessing auditory distance perception using perceptually directed action. In : *Perception & Psychophysics*. septembre 1998. Vol. 60, n° 6, pp. 966-980. DOI 10.3758/BF03211932.

LORD RAYLEIGH, 1907. XII. On our perception of sound direction. In : *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. 1 février 1907. Vol. 13, n° 74, pp. 214-232. DOI 10.1080/14786440709463595.

LOYAU, Fanny, 2007. *Expressions des états mentaux et émotionnels de l'humain en intéraction : ébauches du « Feeling of Thinking »* [en ligne]. These de doctorat. S.l. : Grenoble INPG. [Consulté le 17 juin 2020]. Disponible à l'adresse : http://www.theses.fr/2007INPG0171.

LU, J.M., LU, C., CHEN, Y. et HSU, Y., 2011. TRiCmini - A Telepresence Robot towards Enriched Quality of Life of the Elderly. In : *Proceedings of the Asia Pacific eCare and TeleCare Congress* [en ligne]. S.l. : s.n. 2011. [Consulté le 22 juillet 2019]. Disponible à l'adresse : http://designer.mech.yzu.edu.tw/articlesystem/article/compressedfile/(2011-05-18)%20TRiCmini%20-%20A%20Telepresence%20Robot%20towards%20Enriched%20Quality%20of%20Life%20of%20the%20Elderly.pdf.

LU, Yan, 2015. *Etude contrastive de la prosodie audio-visuelle des affects sociaux en chinois mandarin vs.français : vers une application pour l'apprentissage de la langue étrangère ou seconde* [en ligne]. phdthesis. S.l. : Université Grenoble Alpes. [Consulté le 18 mai 2020]. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-01227267.

LUO, Jinhong, HAGE, Steffen R. et MOSS, Cynthia F., 2018. The Lombard Effect: From Acoustics to Neural Mechanisms. In : *Trends in Neurosciences*. décembre 2018. Vol. 41, n° 12, pp. 938-949. DOI 10.1016/j.tins.2018.07.011.

MAC, Dang Khoa, 2012. *Génération de parole expressive dans le cas des langues à tons* [en ligne]. phdthesis. S.l. : Université de Grenoble. [Consulté le 18 mai 2020]. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-00859201.

MAGNANI, Romain, AUBERGÉ, Véronique, BAYOL, Clarisse et SASA, Yuko, 2017. Bases of Empathic Animism Illusion: audio-visual perception of an object devoted to becoming perceived as a subject for HRI. In : *VIHAR*. Stockholm, Sweden : s.n. 2017.

MAIR, Gordon M., 2007. Towards Transparent Telepresence. In : SHUMAKER, Randall (éd.), *Virtual Reality* [en ligne]. Berlin, Heidelberg : Springer Berlin Heidelberg. pp. 300-309. [Consulté le 4 janvier 2019]. ISBN 978-3-540-73334-8. Disponible à l'adresse : http://link.springer.com/10.1007/978-3-540-73335-5_33.

MALLON, Isabelle, 2010. Le milieu rural isolé isole-t-il les personnes âgées ? In : *Espace populations sociétés. Space populations societies*. 1 avril 2010. n° 2010/1, pp. 109-119. DOI 10.4000/eps.3967.

MANDRAN, Nadine, 2017. Méthode expérimentale. In : *Formation doctorale*. Ecoles doctorales. 2017.

MARXER, Ricard, BARKER, Jon, ALGHAMDI, Najwa et MADDOCK, Steve, 2018. The impact of the Lombard effect on audio and visual speech recognition systems. In : *Speech Communication*. 1 juin 2018. Vol. 100, pp. 58-68. DOI 10.1016/j.specom.2018.04.006.

MCGREGOR, Peter, HORN, Andrew G. et TODD, Melissa A., 1985. Are Familiar Sounds Ranged More Accurately? In : *Perceptual and Motor Skills*. décembre 1985. Vol. 61, n° 3_suppl, pp. 1082-1082. DOI 10.2466/pms.1985.61.3f.1082.

MERSHON, Donald H, 1997. Phenomenal Geometry and the measure-ment of perceived auditory distance. In : *Binaural and Spatial Hearing in Real and Virtual Environments*. Mahwah, New Jersey : Psychology Press. pp. 257-274. ISBN 978-1-317-78026-7.

MIDDLEBROOKS, John C. et GREEN, David M., 1991. Sound Localization by Human Listeners. In : *Annual Review of Psychology*. janvier 1991. Vol. 42, n° 1, pp. 135-159. DOI 10.1146/annurev.ps.42.020191.001031.

MINNAAR, Pauli, OLESEN, S Krarup, CHRISTENSEN, Flemming et MØLLER, Henrik, 2001. Localization with binaural recordings from artificial and human heads. In : *Journal of the Audio Engineering Society*. 2001. Vol. 49, n° 5, pp. 323–336.

MINSKY, Marvin, 1980. Telepresence. In : *s.n.* [en ligne]. 1980. Vol. 2. Disponible à l'adresse : https://spectrum.ieee.org/robotics/artificial-intelligence/telepresence-a-manifesto.

MØLLER, Henrik, 1992. Fundamentals of binaural technology. In : *Applied Acoustics*. 1992. Vol. 36, n° 3-4, pp. 171-218. DOI 10.1016/0003-682X(92)90046-U.

MØLLER, Henrik, HAMMERSHØI, Dorte, BOJE, Clemen et SØRENSEN, Michael Friis, 1999. Evaluation of artificial heads in listening tests. In : *AES*. 1999. Vol. 47, n° 3, pp. 83-100.

MONSON, Brian B., HUNTER, Eric J. et STORY, Brad H., 2012. Horizontal directivity of low- and high-frequency energy in speech and singing. In : *The Journal of the Acoustical Society of America*. juillet 2012. Vol. 132, n° 1, pp. 433-441. DOI 10.1121/1.4725963.

MOORE, Roger K, 2017. Appropriate Voices for Artefacts: Some Key Insights. In : *1st International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*. S.l. : s.n. 2017.

NEUSTAEDTER, Carman, VENOLIA, Gina, PROCYK, Jason et HAWKINS, Daniel, 2016. To Beam or Not to Beam: A Study of Remote Telepresence Attendance at an Academic Conference. In : *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* [en ligne]. New York, NY, USA : ACM. 2016. pp. 418–431. [Consulté le 22 juillet 2019]. ISBN 978-1-4503-3592-8. Disponible à l'adresse : http://doi.acm.org/10.1145/2818048.2819922.

NEWMAN, Rochelle S., 2003. Prosodic differences in mothers' speech to toddlers in quiet and noisy environments. In : *Applied Psycholinguistics*. décembre 2003. Vol. 24, n° 4, pp. 539-560. DOI 10.1017/S0142716403000274.

NILSSON, Niels Christian, NORDAHL, Rolf et SERAFIN, Stefania, 2016. Immersion Revisited: A Review of Existing Definitions of Immersion and Their Relation to Different Theories of Presence. In : *Human Technology* [en ligne]. 2016. Vol. 12. [Consulté le 3 juillet 2020]. DOI 10.17011/ht/urn.201611174652. Disponible à l'adresse : https://jyx.jyu.fi/handle/123456789/52083.

NISHIO, Shuichi, ISHIGURO, Hiroshi et HAGITA, Norihiro, 2007. Geminoid: Teleoperated Android of an Existing Person. In : *Humanoid Robots: New Developments* [en ligne]. 1 juin 2007. [Consulté le 14 janvier 2020]. DOI 10.5772/4876. Disponible à l'adresse : https://www.intechopen.com/books/humanoid_robots_new_developments/geminoid__teleoperated_android_of_an_existing_person.

NOWAK, Kristine, 2001. Defining and differentiating copresence, social presence and presence as transportation. In : *Presence 2001 Conference*. Philadelphia : s.n. 2001. pp. 1–23.

OH-HUN KWON, SEONG-YONG KOO, YOUNG-GEUN KIM et DONG-SOO KWON, 2010. Telepresence robot system for English tutoring. In : *2010 IEEE Workshop on Advanced Robotics and its Social Impacts* [en ligne]. Seoul, Corée du sud : IEEE. octobre 2010. pp. 152-155. [Consulté le 15 janvier 2020]. ISBN 978-1-4244-9122-3. Disponible à l'adresse : http://ieeexplore.ieee.org/document/5679999/.

OKAMURA, A.M., 2004. Methods for haptic feedback in teleoperated robot-assisted surgery. In : *Industrial Robot: An International Journal*. 1 janvier 2004. Vol. 31, n° 6, pp. 499-508. DOI 10.1108/01439910410566362.

OVADIA, Daniela, 2016. Les insoutenables expériences sur l'amour maternel. In : *Pour la Science* [en ligne]. octobre 2016. Vol. Cerveau&Psycho, n° 81. [Consulté le 19 février 2020]. Disponible à l'adresse : https://www.pourlascience.fr/theme/amour-et-couple/les-insoutenables-experiences-sur-lamour-maternel-9251.php.

PAEPCKE, Andreas, SOTO, Bianca, TAKAYAMA, Leila, KOENIG, Frank et GASSEND, Blaise, 2011. Yelling in the hall: using sidetone to address a problem with mobile remote presence systems. In : *Proceedings of the 24th annual ACM symposium on User interface software and technology - UIST '11* [en ligne]. Santa Barbara, California, USA : ACM Press. 2011. pp. 107. [Consulté le 4 janvier 2019]. ISBN 978-1-4503-0716-1. Disponible à l'adresse : http://dl.acm.org/citation.cfm?doid=2047196.2047209.

PAUL, Stephan, 2009. Binaural Recording Technology: A Historical Review and Possible Future Developments. In : *Acta Acustica united with Acustica*. 1 septembre 2009. Vol. 95, n° 5, pp. 767-788. DOI 10.3813/AAA.918208.

PEARSONS, K.S., BENNETT, R.L. et FIDELL, S., 1977. EPA-600/1-77-025 : *Speech levels in various noise environments* [en ligne]. Washington DC. U.S. Environmental Protection Agency. [Consulté le 16 janvier 2019]. Disponible à l'adresse : https://cfpub.epa.gov/si/si_public_record_Report.cfm?Lab=ORD&dirEntryID=45786.

PELEGRÍN-GARCÍA, David, SMITS, Bertrand, BRUNSKOG, Jonas et JEONG, Cheol-Ho, 2011. Vocal effort with changing talker-to-listener distance in different acoustic environments. In : *The Journal of the Acoustical Society of America*. avril 2011. Vol. 129, n° 4, pp. 1981-1990. DOI 10.1121/1.3552881.

PÉPIOT, Erwan, 2013. *Voix de femmes, voix d'hommes: différences acoustiques, identification du genre par la voix et implications psycholinguistiques chez les locuteurs anglophones et francophones* [en ligne]. Linguistique. S.l. : Université Paris VIII Vincennes-Saint Denis. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-00821462.

PETKOVA, Valeria I. et EHRSSON, H. Henrik, 2008. If I Were You: Perceptual Illusion of Body Swapping. In : *PLOS ONE*. 3 décembre 2008. Vol. 3, n° 12, pp. e3832. DOI 10.1371/journal.pone.0003832.

PHILBECK, John W. et MERSHON, Donald H., 2002. Knowledge about typical source output influences perceived auditory distance. In : *The Journal of the Acoustical Society of America*. 2002. Vol. 111, n° 5, pp. 1980. DOI 10.1121/1.1471899.

PICARD, Dominique, 1992. De la communication à l'interaction : l'évolution des modèles. In : *Communication & Langages*. 1992. Vol. 93, n° 1, pp. 69-83. DOI 10.3406/colan.1992.2380.

PICK, Herbert L., SIEGEL, Gerald M., FOX, Paul W., GARBER, Sharon R. et KEARNEY, Joseph K., 1989. Inhibiting the Lombard effect. In : *The Journal of the Acoustical Society of America*. 1 février 1989. Vol. 85, n° 2, pp. 894-900. DOI 10.1121/1.397561.

POČTA, Peter et KOMPERDA, Oliwia, 2016. A Black Box Analysis of WebRTC Mouth-to-Ear Delays. In : *Communications – Scientific Letters of the University of Zilina* [en ligne]. 2016. [Consulté le 23 avril 2019]. Disponible à l'adresse : https://www.academia.edu/22130349/A_Black_Box_Analysis_of_WebRTC_Mouth-to-Ear_Delays.

POLLOW, Martin, 2015. *Directivity patterns for room acoustical measurements and simulations*. Berlin : Logos Verlag Berlin GmbH. Aachener Beiträge zur Technischen Akustik, vol. 22. ISBN 978-3-8325-4090-6. TA365 .P65 2015

POSTMA, Barteld NJ et KATZ, Brian FG, 2016. Dynamic voice directivity in room acoustic auralizations. In : *German Annual Conference on Acoustics (DAGA)*. S.l. : s.n. 2016. pp. 352-355.

PRABLANC, Pierre, 2014. *Conversion de la voix pour un robot accompanateur*. Rapport de stage M1. S.l. Grenoble INP - Phelma.

PUNCH, Jerry, JOSEPH-, Antony et RAKERD, Brad, 2004. Most comfortable and uncomfortable loudness levels: six decades of research. In : *American Journal of Audiology*. décembre 2004. Vol. 13, n° 2, pp. 144-157.

RASCON, Caleb et MEZA, Ivan, 2017. Localization of sound sources in robotics: A review. In : *Robotics and Autonomous Systems*. 1 octobre 2017. Vol. 96, pp. 184-210. DOI 10.1016/j.robot.2017.07.011.

RECANZONE, Gregg H., MAKHAMRA, Samia D. D. R. et GUARD, Darren C., 1998. Comparison of relative and absolute sound localization ability in humans. In : *The Journal of the Acoustical Society of America*. février 1998. Vol. 103, n° 2, pp. 1085-1097. DOI 10.1121/1.421222.

REYNOLDS, Alastair, 2015. *La Terre bleue de nos souvenirs* [en ligne]. Paris, France : Bragelonne. [Consulté le 19 juillet 2019]. Disponible à l'adresse : http://fr.feedbooks.com/item/1273418/la-terre-bleue-de-nos-souvenirs.

RILLIARD, Albert, 2000. *Vers une mesure de l'intelligibilité linguistique de la prosodie : évaluation diagnostique des prosodies synthétique et naturelle* [en ligne]. S.l. : Grenoble INPG. [Consulté le 17 juin 2020]. Disponible à l'adresse : http://www.theses.fr/2000INPG0156.

RIZZOLATTI, Giacomo et CRAIGHERO, Laila, 2004. The Mirror-Neuron System. In : *Annual Review of Neuroscience*. 2004. Vol. 27, n° 1, pp. 169-192. DOI 10.1146/annurev.neuro.27.070203.144230.

ROFFLER, Suzanne K. et BUTLER, Robert A., 1968. Factors That Influence the Localization of Sound in the Vertical Plane. In : *The Journal of the Acoustical Society of America*. juin 1968. Vol. 43, n° 6, pp. 1255-1259. DOI 10.1121/1.1910976.

RUMSEY, Francis, 2002. Spatial Quality Evaluation for Reproduced Sound: Terminology, Meaning, and a Scene-Based Paradigm. In : *Journal of the Audio Engineering Society*. 15 septembre 2002. Vol. 50, n° 9, pp. 651-666.

SAADATIAN, Elham, SAMANI, Hooman, VIKRAM, Anshul, PARSANI, Rahul, TEJADA RODRIGUEZ, Lenis et NAKATSU, Ryohei, 2013. Personalizable embodied telepresence system for remote interpersonal communication. In : *2013 IEEE RO-MAN* [en ligne]. Gyeongju : IEEE. août 2013. pp. 226-231. [Consulté le 22 juillet 2019]. ISBN 978-1-4799-0509-6. Disponible à l'adresse : http://ieeexplore.ieee.org/document/6628450/.

SAITO, Tamaki et ANGLES, Jeffrey, 2013. Hikikomori: Adolescence Without End. In : *All Books and Monographs by WMU Authors* [en ligne]. 1 janvier 2013. Disponible à l'adresse : https://scholarworks.wmich.edu/books/31.

SASA, Yuko, 2018. *Intelligence Socio-Affective pour un Robot : primitives langagières pour une interaction évolutive d'un robot de l'habitat intelligent (Socio-affective Intelligence for a Robot: language primitives for evolving interaction with a robot companion in smart home)* [en ligne]. Intelligence Artificielle. S.l. : Université Grenoble Alpes. [Consulté le 5 janvier 2019]. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-01945238/document.

SATO, Hayato, SATO, Hiroshi, MORIMOTO, Masayuki et OTA, Ryo, 2007. Acceptable range of speech level for both young and aged listeners in reverberant and quiet sound fields. In : *The Journal of the Acoustical Society of America*. 1 septembre 2007. Vol. 122, n° 3, pp. 1616-1623. DOI 10.1121/1.2766780.

SCHAFER, R Murray, 1977. *The tuning of the world*. S.l. : Alfred A. Knopf.

SCHENKMAN, Bo N et NILSSON, Mats E, 2010. Human Echolocation: Blind and Sighted Persons' Ability to Detect Sounds Recorded in the Presence of a Reflecting Object. In : *Perception*. avril 2010. Vol. 39, n° 4, pp. 483-501. DOI 10.1068/p6473.

SCHERER, Klaus R., 1995. Expression of emotion in voice and music. In : *Journal of Voice*. 1 septembre 1995. Vol. 9, n° 3, pp. 235-248. DOI 10.1016/S0892-1997(05)80231-0.

SCHNEIDERMANN, Daniel et GRAMAGLIA, Juliette, 2020. 5G : « A-t-on besoin qu'une vache soit connectée en permanence ? » - Par La rédaction | Arrêt sur images. In : *Arrêt sur images* [en ligne]. 7 février 2020. [Consulté le 3 mars 2020]. Disponible à l'adresse : https://www.arretsurimages.net/emissions/arret-sur-images/5g-a-t-on-besoin-quune-vache-soit-connectee-en-permanence.

SHANNON, C E, 1948. A Mathematical Theory of Communication. In : *The Bell system technical journal*. 1948. Vol. 27, n° 3, pp. 379-423.

SHINN-CUNNINGHAM, Barbara, 2003. Acoustics and perception of sound in everyday environments. In : *Proceedings of 3rd International Workshop on Spatial Media* [en ligne]. Aizy-Wakamatsu, Japan : s.n. mars 2003. pp. R442-R444. [Consulté le 27 janvier 2020]. Disponible à l'adresse : https://linkinghub.elsevier.com/retrieve/pii/S0960982215002298.

SHOCHI, Takaaki, 2008. *Prosodie des affects socioculturels en japonais, et anglais: à la recherche des vrais et faux-amis pour le parcours de l'apprenant* [en ligne]. phdthesis. S.l. : Université Stendhal - Grenoble III. [Consulté le 17 juin 2020]. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-00366612.

SHOCHI, Takaaki, RILLIARD, Albert, AUBERGÉ, Véronique et ERICKSON, Donna, 2009. Intercultural Perception of English, French and Japanese Social Affective Prosody. In : *The Role of Prosody in Affective Speech*. Sylvie Hancil. S.l. : Peter Lang. pp. 31-60. ISBN 978-3-03911-696-6.

STEENEKEN, Herman J. M. et HOUTGAST, Tammo, 2014. Basics of the STI-measuring method. In : [en ligne]. 2014. [Consulté le 29 juillet 2019]. Disponible à l'adresse : https://core.ac.uk/display/100729420.

STEUER, Jonathan, 1992. Defining Virtual Reality: Dimensions Determining Telepresence. In : *Journal of Communication*. décembre 1992. Vol. 42, n° 4, pp. 73-93. DOI 10.1111/j.1460-2466.1992.tb00812.x.

STOLL, Brett, REIG, Samantha, HE, Lucy, KAPLAN, Ian, JUNG, Malte F. et FUSSELL, Susan R., 2018. Wait, Can You Move the Robot?: Examining Telepresence Robot Use in Collaborative Teams. In : *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* [en ligne]. New York, NY, USA : ACM. 2018. pp. 14–22. [Consulté le 7 juin 2019]. ISBN 978-1-4503-4953-6. Disponible à l'adresse : http://doi.acm.org/10.1145/3171221.3171243.

SUMMERS, W. Van, PISONI, David B., BERNACKI, Robert H., PEDLOW, Robert I. et STOKES, Michael A., 1988. Effects of noise on speech production: Acoustic and perceptual analyses. In : *The Journal of the Acoustical Society of America*. 1 septembre 1988. Vol. 84, n° 3, pp. 917-928. DOI 10.1121/1.396660.

SZEKELY, Eva, KEANE, Mark T et CARSON-BERNDSEN, Julie, 2015. The effect of soft, modal and loud voice levels on entrainment in noisy conditions. In : *Interspeech*. S.l. : s.n. 2015.

TANAKA, Fumihide, TAKAHASHI, Toshimitsu, MATSUZOE, Shizuko, TAZAWA, Nao et MORITA, Masahiko, 2014. Telepresence robot helps children in communicating with teachers who speak a different language. In : *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction - HRI '14* [en ligne]. Bielefeld, Germany : ACM Press. 2014. pp. 399-406. [Consulté le 15 janvier 2020]. ISBN 978-1-4503-2658-2. Disponible à l'adresse : http://dl.acm.org/citation.cfm?doid=2559636.2559654.

TARTTER, Vivien C., GOMES, Hilary et LITWIN, Elissa, 1993. Some acoustic effects of listening to noise on speech production. In : *The Journal of the Acoustical Society of America*. 1 octobre 1993. Vol. 94, n° 4, pp. 2437-2440. DOI 10.1121/1.408234.

TAVARES, Aida Isabel, 2017. Telework and health effects review. In : *International Journal of Healthcare*. 11 juillet 2017. Vol. 3, n° 2, pp. 30. DOI 10.5430/ijh.v3n2p30.

TERRIEN, Soizic, 2014. *Instruments de la famille des flûtes: analyse des transitions entre régimes* [en ligne]. Acoustique. S.l. : Université Aix-Marseille. Disponible à l'adresse : tel-01142359.

TOBITA, Hiroaki, MARUYAMA, Shigeaki et KUZI, Takuya, 2011. Floating avatar: telepresence system using blimps for communication and entertainment. In : *CHI'11 Extended Abstracts on Human Factors in Computing Systems*. Vancouver, Canada : ACM. 2011. pp. 541–550.

TOSHIMA, Iwaki, AOKI, Shigeaki et HIRAHARA, Tatsuya, 2008. Sound Localization Using an Acoustical Telepresence Robot: TeleHead II. In : *Presence: Teleoperators and Virtual Environments*. 16 juillet 2008. Vol. 17, n° 4, pp. 392-404. DOI 10.1162/pres.17.4.392.

TRAUNMÜLLER, Hartmut, 1994. Conventional, Biological and Environmental Factors in Speech Communication: A Modulation Theory. In : *Phonetica*. 1994. Vol. 51, n° 1-3, pp. 170-183. DOI 10.1159/000261968.

TUBACH, J.P., 1989. Description acoustique. In : *La Parole et son traitement automatique*. Paris; Milan; Barcelone : Masson. pp. 79-130. ISBN 978-2-225-81516-4.

TUFTS, Jennifer B. et FRANK, Tom, 2003. Speech production in noise with and without hearing protection. In : *The Journal of the Acoustical Society of America*. août 2003. Vol. 114, n° 2, pp. 1069-1080. DOI 10.1121/1.1592165.

TURK, Oytun, SCHRÖDER, Marc, BOZKURT, Baris et ARSLAN, Levent M, 2005. Voice Quality Interpolation for Emotional Text-To-Speech Synthesis. In : *INTERSPEECH*. Lisbon, Portugal : s.n. 2005. pp. 797-800.

VAN HEUSDEN, E., PLOMP, R. et POLS, L. C. W., 1979. Effect of ambient noise on the vocal output and the preferred listening level of conversational speech. In : *Applied Acoustics*. 1 janvier 1979. Vol. 12, n° 1, pp. 31-43. DOI 10.1016/0003-682X(79)90037-9.

VANPÉ, Anne, 2011. *Expressions et micro-expressions spontanées de la face et de la voix en Interaction Homme-Machine : esquisse d'un modèle du « Feeling of Thinking ».* [en ligne]. phdthesis. S.l. : Université Stendhal - Grenoble III. [Consulté le 17 juin 2020]. Disponible à l'adresse : https://tel.archives-ouvertes.fr/tel-00625714.

VERTEGAAL, Roel, WEEVERS, Ivo, SOHN, Changuk et CHEUNG, Chris, 2003. GAZE-2: conveying eye contact in group video conferencing using eye-controlled camera direction. In : *Proceedings of the conference on Human factors in computing systems - CHI '03* [en ligne]. Ft. Lauderdale, Florida, USA : ACM Press. 2003. pp. 521. [Consulté le 22 juillet 2019]. ISBN 978-1-58113-630-2. Disponible à l'adresse : http://portal.acm.org/citation.cfm?doid=642611.642702.

WALTERS, M. L., SYRDAL, D. S., KOAY, K. L., DAUTENHAHN, K. et TE BOEKHORST, R., 2008. Human approach distances to a mechanical-looking robot with different robot voice styles. In : *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication* [en ligne]. Munich, Germany : IEEE. août 2008. pp. 707-712. [Consulté le 21 juillet 2020]. ISBN 978-1-4244-2212-8. Disponible à l'adresse : http://ieeexplore.ieee.org/document/4600750/.

WATZLAWICK, Paul, BEAVIN, Janet Helmick, JACKSON, Don D et MORCHE, J., 1972. *Une logique de la communication*. Points. S.l. : s.n. Points Essais.

WIENER, Norbert, 1948. *Cybernetics Or Control and Communication in the Animal and the Machine*. MIT Press. Cambridge, Massachusetts : s.n. ISBN 978-0-262-73009-9.

WIGHTMAN, Frederic L. Kistler, 1990. *Hearing in three dimensions: Sound localization* [en ligne]. S.l. [Consulté le 3 janvier 2019]. Disponible à l'adresse : https://ntrs.nasa.gov/search.jsp?R=19910004658.

WINKIN, Yves, 1981. *La Nouvelle Communication, Yves Winkin* [en ligne]. S.l. : s.n. [Consulté le 6 février 2019]. Points Essais. Disponible à l'adresse : http://www.seuil.com/ouvrage/la-nouvelle-communication-yves-winkin/9782757844465.

WINKWORTH ALISON L. et DAVIS PAMELA J., 1997. Speech Breathing and the Lombard Effect. In : *Journal of Speech, Language, and Hearing Research*. 1 février 1997. Vol. 40, n° 1, pp. 159-169. DOI 10.1044/jslhr.4001.159.

WINTER, Bodo, 2013. Linear models and linear mixed effects models in R with linguistic applications. In : *arXiv:1308.5499 [cs]* [en ligne]. 26 août 2013. [Consulté le 10 janvier 2019]. Disponible à l'adresse : http://arxiv.org/abs/1308.5499.

WISNIEWSKI, Matthew G., III, Eduardo Mercado, GRAMANN, Klaus et MAKEIG, Scott, 2012. Familiarity with Speech Affects Cortical Processing of Auditory Distance Cues and Increases Acuity. In : *PLOS ONE*. 20 juillet 2012. Vol. 7, n° 7, pp. e41025. DOI 10.1371/journal.pone.0041025.

YADAV, Manuj, CABRERA, Densil A., MIRANDA, Luis, MARTENS, William L., LEE, Doheon et COLLINS, Ralph, 2013. Investigating Auditory Room Size Perception with Autophonic Stimuli. In : *Audio Engineering Society Convention 135* [en ligne]. S.l. : Audio Engineering Society. 16 octobre 2013. [Consulté le 27 janvier 2020]. Disponible à l'adresse : http://www.aes.org/e-lib/browse.cfm?elib=16984.

YAIRI, Satoshi, IWAYA, Yukio et SUZUKI, Yôiti, 2007. Estimation of detection threshold of system latency of virtual auditory display. In : *Applied Acoustics*. 1 août 2007. Vol. 68, n° 8, pp. 851-863. DOI 10.1016/j.apacoust.2006.12.005.

YANG, Lillian, JONES, Brennan, NEUSTAEDTER, Carman et SINGHAL, Samarth, 2018. Shopping Over Distance through a Telepresence Robot. In : *Proceedings of the ACM on Human-Computer Interaction*. 1 novembre 2018. Vol. 2, n° CSCW, pp. 1-18. DOI 10.1145/3274460.

ZAHORIK, Pavel, 2005. Auditory Distance Perception in Humans: A Summary of Past and Present Research. In : *ACTA ACUSTICA UNITED WITH ACUSTICA*. 2005. Vol. 91, pp. 12.

ZAHORIK, Pavel, BRUNGART, Douglas S et BRONKHORST, Adelbert W, 2005. Auditory distance perception in humans: A summary of past and present research. In : *ACTA Acustica united with Acustica*. 2005. Vol. 91, n° 3, pp. 409–420.

ZOLLINGER, Sue Anne et BRUMM, Henrik, 2011. The evolution of the Lombard effect: 100 years of psychoacoustic research. In : *Behaviour*. 2011. Vol. 148, n° 11-13, pp. 1173-1198. DOI 10.1163/000579511X605759.

PUBLICATIONS

DAVAT, Ambre, 2015. *Développement d'une application temps réel de conversion de voix pour un robot compagnon*. Rapport de stage M1. S.l. Grenoble INP.

DAVAT, Ambre, AUBERGÉ, Véronique et FENG, Gang, 2018a. Integrating Socio-Affective Information in Physical Perception aimed to Telepresence Robots. In : *2018 International Conference on Behavioral, Economic and Socio-cultural Computing (BESC)*. Kaohsiung, Taiwan : s.n. 12 novembre 2018.

DAVAT, Ambre, AUBERGÉ, Véronique et FENG, Gang, 2018b. Vers un modèle du « toucher vocal » pour la communication ubiquïte. In : *XXXIIe Journées d'Études sur la Parole* [en ligne]. S.l. : ISCA. 4 juin 2018. p. 303-311. [Consulté le 24 avril 2019]. Disponible à l'adresse : http://www.isca-speech.org/archive/JEP_2018/abstracts/192737.html.

DAVAT, Ambre, FENG, Gang et AUBERGÉ, Véronique, 2019. A Study on the Lombard Effect in Telepresence Robotics. In : *Proceedings of the 2nd International Workshop on Vocal Interactivity in-and-between Humans, Animals and Robots*. Londres : s.n. 2019.

DAVAT, Ambre, QUERCIA, Alain, AUBERGÉ, Véronique, BOREL, Natacha et FENG, Gang, 2019. Co-creation of a Transitional Smart Sculpture for Voice Changes. In : *Proceedings of the Thirteenth International Conference on Tangible, Embedded, and Embodied Interaction* [en ligne]. New York, NY, USA : ACM. 2019. p. 97–104. [Consulté le 24 avril 2019]. Disponible à l'adresse : http://doi.acm.org/10.1145/3294109.3295654.

DAVAT, Ambre, AUBERGÉ, Véronique et FENG, Gang, 2020. Can we hear physical and social space together through prosody? In : *10th International Conference on Speech Prosody 2020* [en ligne]. S.l. : ISCA. 25 mai 2020. p. 715-719. [Consulté le 6 juillet 2020]. Disponible à l'adresse : http://www.isca-speech.org/archive/SpeechProsody_2020/abstracts/42.html.

DAVAT, Ambre, QUERCIA, Alain et AUBERGÉ, Véronique, 2020. Aporia : Une performance pluri-disciplinaire Arts et IA. In : *Journées d'Informatique Théâtrale* [en ligne]. Grenoble, France : s.n. février 2020. [Consulté le 22 juillet 2020]. Disponible à l'adresse : https://hal.archives-ouvertes.fr/hal-02899525.

LOGITHÈQUE

Analyse des données

AUDACITY TEAM, 2020. *Audacity(R): Free Audio Editor and Recorder [Computer application]* [en ligne]. S.l. : s.n. Disponible à l'adresse : https://audacityteam.org/.

BATES, Douglas et MÄCHLER, Martin, 2015. Fitting Linear Mixed-Effects Models Using {lme4}. In : *Journal of Statistical Software*. 2015. Vol. 67, n° 1, p. 1-48. DOI 10.18637/jss.v067.i01.

BOERSMA, Paul et WEENINK, David, 2019. *PRAAT: doing phonetics by computer* [en ligne]. S.l. : s.n. [Consulté le 22 janvier 2019]. Disponible à l'adresse : http://www.praat.org/.

MATLAB, 2016. *Matlab*. Natick, Massachusetts : The MathWorks Inc.

R CORE TEAM, 2020. *R: A Language and Environment for Statistical Computing* [en ligne]. Vienne, Autriche : R Foundation for Statistical Computing. Disponible à l'adresse : https://www.R-project.org/.

Sculpture connectée pour le spectacle Aporia

BLENDER ONLINE COMMUNITY, 2016. *Blender - a 3D modelling and rendering package* [en ligne]. Stichting Blender Foundation, Amsterdam : Blender Foundation. Disponible à l'adresse : http://www.blender.org.

CREATIVE TECHNOLOGY, 2010. *OpenAL* [en ligne]. S.l. : s.n. Disponible à l'adresse : www.openal.org.

MAZZONI, Dominic, 2013. *libresample : Real-time library for sample rate conversion* [en ligne]. S.l. : s.n. Disponible à l'adresse : http://github.com/minorninth/libresample.

POETTERING, Lennart, KING, Shahms E. King, KASKINEN, Tanu, GUTHRIE, Colin, RAGHAVAN, Arun et HENNINGSSON, David, 2019. *PulseAudio API* [en ligne]. S.l. : s.n. Disponible à l'adresse : https://freedesktop.org/software/pulseaudio/doxygen/index.html.

Robot de téléprésence

FABMSTIC, 2020. *fabMSTICLig/RobAIR on GitHub* [en ligne]. Grenoble, France : s.n. [Consulté le 12 avril 2019]. Disponible à l'adresse : https://github.com/fabMSTICLig/RobAIR.

ANNEXE A :
CARACTÉRISTIQUES DES ROBOTS DE TÉLÉPRÉSENCE

Cette annexe contient des tableaux récapitulant les caractéristiques techniques des robots de téléprésence disponibles actuellement dans le commerce. Elles sont classées en trois catégories : toucher vocal, toucher visuel et navigation.

	Toucher vocal			
Robot	**Microphones**	**Annulation de bruit**	**Haut-parleurs**	**Réglage du volume**
Ava 500	NA	NA	NA	boutons sur le robot
Beam+	4	oui	NA	automatique
BeamPro	6	oui	NA	automatique
BotEyes	variable	NA	variable	NA
Carl	NA	oui	amplification	NA
Collaborate i/o	variable	NA	variable	NA
Double	6	oui	amplification	NA
Endurance	NA	NA	variable	NA
Giraff	NA	NA	NA	NA
Jazz	1	NA	NA	NA
Kubi	variable	NA	variable	NA
Ohmni	1 (omnidirectionnel, 100Hz - 10kHz)	oui	250Hz - 14 kHz jusqu'à 48dB	oui
Padbot U1	variable	NA	variable	NA
Padbot T1	variable	NA	variable	NA
PadBotP2	plusieurs	NA	2 woofers	NA
PeopleBot	NA	NA	NA	NA
QB	1	NA	1	NA
RP2W	NA	NA	NA	NA
SelfieBot	variable	NA	variable	NA
Swivl	jusqu'à 5 (portés par les interlocuteurs)	oui	variable	NA
Synergy Swan	variable	NA	variable	NA
Tabletop TeleMe	variable	NA	variable	NA
TeleMe	variable	NA	variable	NA
Teleporter	NA	NA	NA	NA
Twinbot	NA	NA	NA	NA
Ubbo	3	NA	NA	NA
VGo	4	oui	1 haut-parleur + 1 woofer	boutons sur le robot
Vita	mono, directionnel, 50Hz - 19kHz	NA	2 hauts-parleurs + 1 woofer (jusqu'à 100 dB)	NA
Webot	1	NA	NA	NA

	Toucher visuel		Navigation	
Robot	**Caméra**	**Taille de l'écran (cm)**	**Détection d'obstacles**	**Parking automatique**
Ava 500	NA	55 cm	oui	oui
Beam+	2 caméras grand angle	26 cm	NA	oui
BeamPro	2 caméras grand angle (zoom x3)	43 cm	NA	oui
BotEyes	variable	variable	NA	NA
Carl	webcam HD	43 cm	oui	oui
Collaborate i/o	zoom x18	variable	inutile (base fixe)	inutile (base fixe)
Double	2 caméras (1 grand angle + 1 zoom)	25 cm	oui	oui
Endurance	NA	18 cm	oui	NA
Giraff	NA	15 cm	NA	oui
Jazz	1	13 cm ?	oui	oui
Kubi	variable	variable	inutile (base fixe)	inutile (base fixe)
Ohmni	2 caméras grand angle	27 cm	NA	oui
Padbot U1	variable	variable	oui	oui
Padbot T1	variable	variable	oui	NA
PadBotP2	NA	25 cm	oui	NA
PeopleBot	NA	20 cm ?	oui	NA
QB	2 caméras	10 cm ?	oui	NA
RP2W	1 caméra orientable	17 cm ?	oui	NA
SelfieBot	variable	variable	inutile (base fixe)	inutile (base fixe)
Swivl	variable	variable	inutile (base fixe)	inutile (base fixe)
Synergy Swan	variable	variable	NA	NA
Tabletop TeleMe	variable	variable	inutile (base fixe)	inutile (base fixe)
TeleMe	variable	variable	oui	NA
Teleporter	caméra HD	51 cm	oui	oui
Twinbot	NA	25 cm ?	NA	NA
Ubbo	2 caméras	34 cm	oui	NA
VGo	1 zoom (x5) inclinable	15 cm	oui	oui
Vita	2 caméras (zoom)	tête : 38 cm poitrine : 22 cm ?	oui	oui
Webot	1	20 cm ?	oui	NA

Légende :

Information claire

Information non pertinente (inutile, ou qui dépend du modèle de tablette / smartphone utilisé)

Information mal renseignée

? Taille de la diagonale estimée à partir des photos du robot

ANNEXE B :
MESURES DE L'INTENSITÉ SONORE

Il existe un grand nombre de manières de mesurer l'intensité sonore. Cette annexe a été rédigée pour faire le point sur les différentes mesures rencontrées au cours de la thèse, et présenter celle que nous avons choisie pour analyser nos enregistrements.

Intensité acoustique

Une **source sonore** est un objet qui vibre mécaniquement dans un fluide. Cette vibration produit une variation de pression qui se propage sans transport de matière, mais par succession de compression / dilatation des molécules du fluide. Le son est donc une onde, sujette aux mêmes phénomènes que la lumière ou les vagues : réflexion, diffraction, diminution de l'intensité avec la distance, interférence avec les autres ondes sonores...

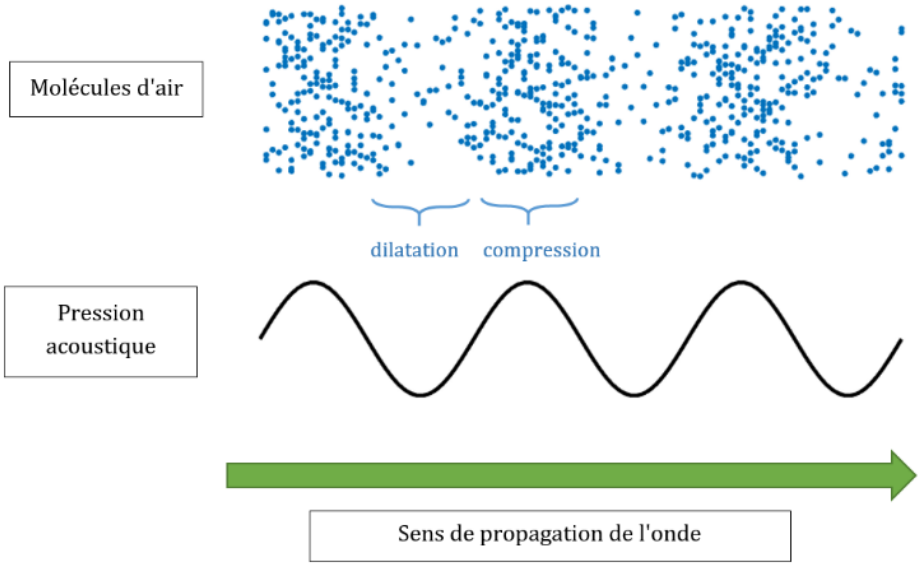


Figure 132 : Propagation d'une onde acoustique dans l'air

La **pression acoustique** p est tout simplement la pression de l'air en un point. Elle oscille autour de la pression atmosphérique (environ 1013 hPa au niveau de la mer).

Un sonomètre permet de mesurer le **niveau de pression acoustique**, défini par :

$$L_p = 20 \log_{10}\left(\frac{p_{eff}}{p_{ref}}\right)$$ unité : **dB SPL** (Sound Pressure Level)

$p_{eff} = \sqrt{\frac{1}{T}\int_{t_0}^{t_0+T} p^2(t)dt}$:	valeur efficace de la pression acoustique calculée sur une fenêtre de durée T
$p_{ref} = 20\ \mu Pa$:	pression de référence, définie de sorte que 0 dB corresponde au seuil d'audibilité.

Par ailleurs, le **niveau d'intensité acoustique** est défini par :

$$L_i = 10 \log_{10}\left(\frac{I}{I_{ref}}\right) \qquad \text{unité : } \textbf{dB SPL} \text{ (Sound Pressure Level)}$$

I :	intensité acoustique
I_{ref} :	définie de sorte que $L_i = L_p$ sous certaines conditions, le plus souvent vérifiées[28]

Tandis que la pression acoustique ne concerne qu'un point de l'espace, l'**intensité acoustique** est un vecteur de puissance acoustique par unité de surface, défini par :

$$\vec{I}(\vec{x}, t) = \frac{1}{T}\int_{t_0}^{t_0+T} p.\vec{v}\, dt$$

La notion d'intensité acoustique permet de décrire la manière dont se répartit la **puissance acoustique** d'une source sonore P. Si cette puissance se répartit de la même manière dans toutes les directions de l'espace, la source est qualifiée d'**omnidirectionnelle**, et dans ce cas, le niveau d'intensité acoustique à une distance r de la source peut être calculée simplement par :

$$I = \frac{P}{4\pi r^2} \qquad \text{unité : } \textbf{W/m²}$$

Notons que la puissance acoustique P ne peut pas être mesurée directement, mais doit être déduite à partir d'une mesure du niveau d'intensité acoustique I à une distance donnée. Par convention, cette mesure s'effectue généralement à 1 m de la source sonore.

Si les notions précédentes permettent de décrire l'intensité sonore, elles ne sont pas représentatives de l'intensité perçue réellement par l'auditeur. Tout d'abord, le champ auditif humain est limité : nous ne pouvons pas percevoir les infrasons (< 20 Hz) et les ultrasons (> 20 kHz), bien que d'autres espèces animales soient capables de les entendre. De plus, la sensibilité de l'oreille humaine varie en fonction de la fréquence : par exemple, pour un même niveau d'intensité acoustique, les sons graves de fréquence inférieure à 100 Hz sont perçus moins fort que les sons aigus. Il a donc fallu inventer une autre notion pour pouvoir décrire le niveau d'intensité acoustique perçu par un auditeur : il s'agit de la **sonie**, définie expérimentalement à l'aide de tests perceptifs. Son unité n'est pas le dB SPL, mais le phone : deux sons purs[29] ont la même valeur en phones, s'ils paraissent de même intensité à l'auditeur. La valeur du phone est indexée de sorte que pour un son pur de 1000 Hz la mesure en phone soit égale à la mesure en dB SPL. Pour connaître la sonie d'un son pur, on utilise directement les courbes isosoniques, représentées en Figure 133.

---

[28] Les ondes acoustiques produites par une source qui émet dans toutes les directions sont sphériques, mais peuvent être assimilées à des ondes planes à une distance suffisante de la source grâce à un développement limité. Il est alors possible d'exprimer l'intensité acoustique en fonction de la pression acoustique : $I = \frac{P^2}{\rho c}$ avec ρ la masse volumique de l'air et c la célérité du son (environ 340 m/s)

[29] Un son pur est un signal sinusoïdal de fréquence et d'amplitude maximale constantes au cours du temps.

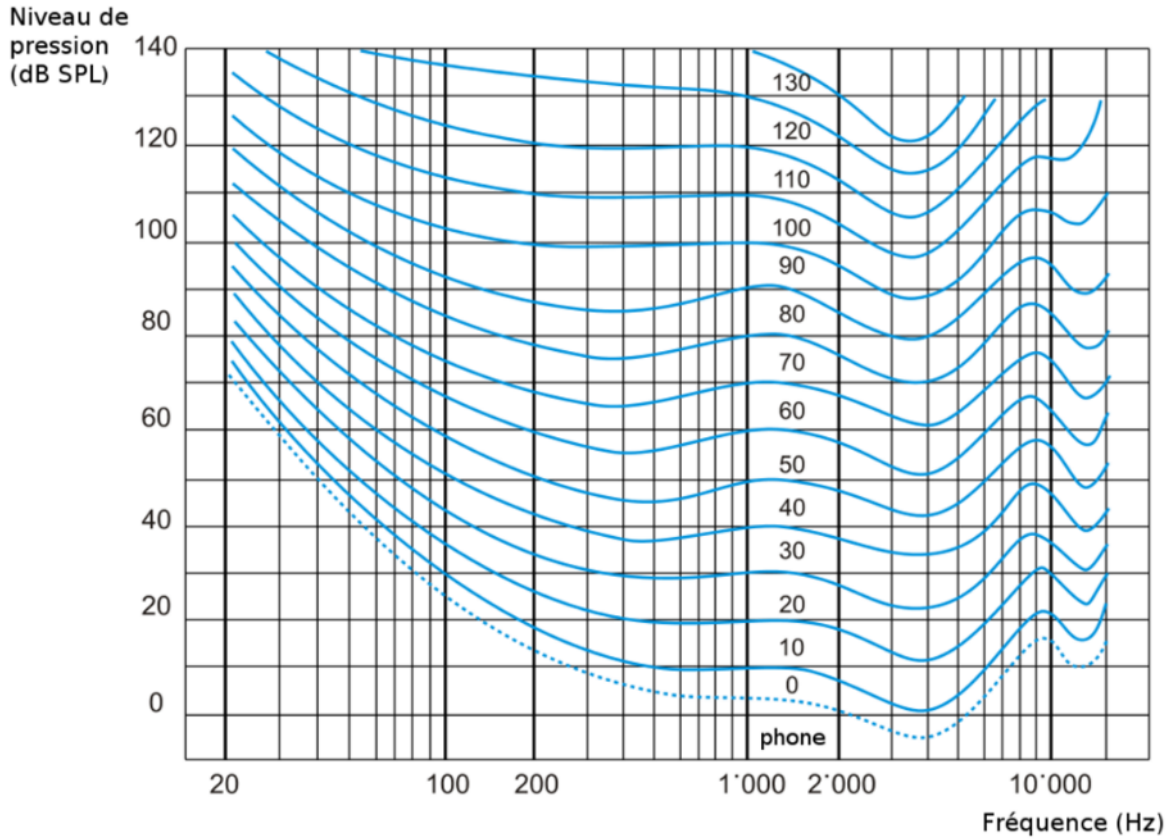


Figure 133 : Courbes isosoniques selon la norme ISO 226:2003
(auditeurs normo-entendant âgés de 18 à 25 ans)

En revanche, pour estimer la sonie d'un son complexe, on utilise des courbes de pondération en fréquence, basées sur les courbes isosoniques. Les principales sont représentées en Figure 134. Elles font l'objet de normes et sont notamment utilisées pour la mesure de bruits ambiants. La courbe de pondération A a pour unité associé le dB A et permet d'estimer la sonie de sons d'intensité faible, puisqu'elle est basée sur la courbe isosonique de 40 phones. C'est la courbe de pondération la plus utilisée, car elle est notamment citée dans les textes législatifs. Pourtant, il est préférable de faire des mesures en dB B pour des sons d'intensité modérée (70 phones) et en dB C pour des sons d'intensité élevée (100 phones), car la pondération A sous-estime l'importance des basses fréquences. En pratique, un sonomètre sophistiqué permet d'effectuer différentes mesures du niveau d'intensité acoustique : dB SPL, dB A, dB B, dB C…

En audiométrie, il existe également d'autres mesures : le **dB HTL** (Hearing Threshold Level) ou **dB HL** (Hearing Level), qui servent à mesurer des pertes d'audibilité. Ainsi une personne jeune et normo-entendante présente un audiogramme plat de 0 dB HL. Avec l'âge le seuil d'audibilité diminue, en particulier dans les hautes fréquences.

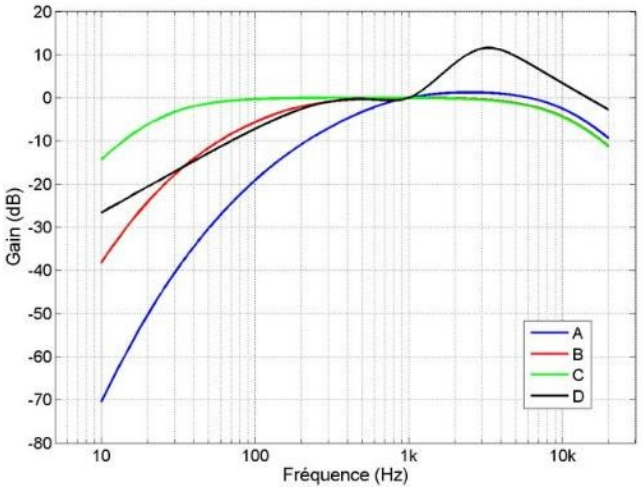


Figure 134 : Courbes de pondération fréquentielle A, B, C et D. (Minard, 2013)

Intensité électrique

Lorsque le signal acoustique est converti en signal électrique, l'information de son niveau d'intensité acoustique est perdue. En effet, l'intensité du signal électrique dépend directement de la sensibilité du capteur utilisé, pas seulement du niveau d'intensité acoustique ; donc à moins de disposer d'un capteur soigneusement calibré, il est impossible de deviner le niveau d'intensité initial. Cependant, il existe des mesures de niveau d'intensité électrique, afin de pouvoir comparer deux signaux électriques entre eux :

$$L_v = 20 \log_{10}\left(\frac{V_{eff}}{V_{ref}}\right)$$

$V_{eff} = \sqrt{\frac{1}{T}\int_{t_0}^{t_0+T} V^2(t)dt}$:	valeur efficace du signal électrique sur une fenêtre temporelle de durée T
V_{ref} :	tension de référence

L'unité de cette mesure dépend de la tension de référence utilisée :

Tension de référence V_{ref}	**Unité de L_v**
0,775 V	dBu
1 V	dBV

Intensité numérique

Enfin, le signal électrique peut être numérisé grâce à un convertisseur analogique-numérique, donc il existe des mesures numériques de l'intensité. La plus simple est définie par :

$$L_{FS} = 20 \log_{10}\left(\frac{y_{eff}}{y_{max}}\right)$$ unité : **dB FS** (*Full-scale* : pleine échelle)

$y_{eff} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\lvert y_i\rvert^2}$:	valeur efficace du signal numérique sur N échantillons
y_{max} :	valeur la plus grande pouvant être représentée par l'échelle numérique (centrée en 0)

C'est ce type de mesure qui est utilisée par le logiciel Praat pour mesurer l'intensité d'un signal sonore ; la valeur de référence y_{max} étant fixée arbitrairement à 20 µPa, soit le seuil d'audibilité. Ainsi, si y était une mesure numérique de la pression acoustique, la mesure fournit par Praat serait égale au niveau d'intensité acoustique (dB SPL).

De la même manière que le dB SPL ne tient pas compte de la sensibilité humaine, les mesures en dBu, dBV et dB FS ne permettent pas d'évaluer la manière dont le signal sonore sera perçu par un auditeur. Une autre métrique a donc dû être développée, notamment en réponse à la guerre du volume[30], afin de permettre aux diffuseurs de normaliser leurs contenus.

Cette métrique est définie par les normes UIT BS. 1770 et EBU R128 et a pour unité le **LUFS** (*Loudness Unit Full Scale*, soit unité de sonie relativement à la pleine échelle numérique). Elle fait appel à une courbe de pondération K, afin d'évaluer de manière pertinente la sonie d'un signal audio numérique. Cette pondération K est obtenue à partir de deux filtres numériques : un pré-filtre, qui modélise les effets acoustiques de la tête en rehaussant les hautes fréquences, et un filtre passe-haut dérivé de la pondération B, qui modélise la sensibilité de l'oreille humaine. La courbe de pondération K est représentée en Figure 135. Outre cette pondération, la norme prévoit un système de seuillage, pour que les parties les plus calmes du signal audio n'impactent pas la mesure globale.

[30] Avant 2006, il n'y avait pas de norme sur le niveau de sonie des programmes audio-visuels. Certains créateurs de contenu audio, en particulier les publicistes, choisissaient donc de mixer leurs programmes de manière à en augmenter artificiellement la sonie. Ces pratiques ont été qualifiées de guerre du volume (*Loudness war*), chaque annonceur s'efforçant de sonner plus fort que ses concurrents, en dépit des distorsions causées au signal audio. L'augmentation des niveaux de sonie par les studios d'enregistrement traduit également une évolution des conditions d'écoute. En effet, il n'est pas possible de percevoir les sons faibles dans un environnement bruyant, comme une voiture ou un transport en commun ; or ce type d'écoute c'est généralisé depuis l'invention du baladeur dans les années 80 (Marie Georgescu de Hillerin, 2015).

Notons que toutes ces mesures d'intensité n'ont pas le même ordre de grandeur. Ainsi, pour des mesures acoustiques, la valeur de référence avancée est généralement de 60 dB (SPL, A, B, C ou D) pour une conversation dans une pièce calme, tandis que le niveau de référence conseillé pour les contenus audio-visuels est de - 23 LUFS.

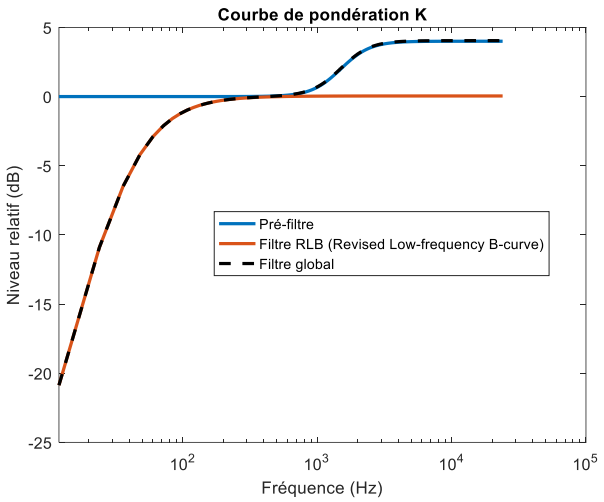
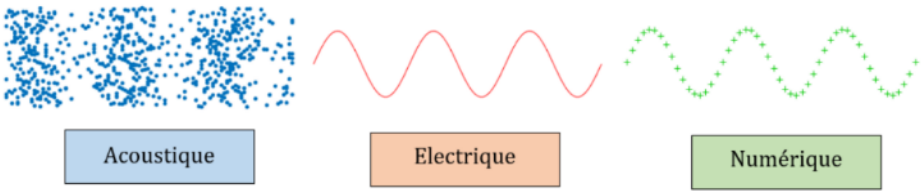


Figure 135 : Courbe de pondération K,
obtenu à partir de deux filtres numériques définis par la norme UIT BS. 1770

Récapitulatif des principales mesures de l'intensité

Il existe donc trois domaines permettant d'étudier et de manipuler les signaux sonores : l'acoustique, l'électronique, et le numérique. Des mesures d'intensité ont donc été définies pour chaque discipline. Par ailleurs, certaines mesures ont été conçues spécifiquement pour estimer la sonie, c'est-à-dire la sensation auditive produite chez un être humain.



	Acoustique	Electrique	Numérique
Mesure du niveau d'intensité	dB(SPL)	dB(V), dB(u)	dB(Praat), dB(FS)
Mesure de la sonie	dB(A), dB(B), dB(C)		LUFS

Figure 136 : Récapitulatif des principales mesures de l'intensité sonore

Il n'est pas possible de retrouver les mesures du niveau d'intensité acoustique à partir des mesures d'intensité électrique ou numérique, à moins d'avoir des informations très précises concernant les conditions d'enregistrement. En effet, le niveau d'intensité électrique dépend initialement de la sensibilité du microphone utilisé et de la distance à la source sonore. Ce signal de quelques millivolts (niveau MIC) est ensuite amplifié jusqu'à environ 1 V (niveau LINE),

niveau standard qui permet l'interconnexion avec d'autres appareils audio (Boller, 2016). Il est également important d'amplifier le signal avant une conversion analogique-numérique, car celle-ci ajoute un bruit de conversion. En conséquence, l'intensité d'un signal de parole numérique ne dit rien de l'intensité produite par le locuteur au moment de l'enregistrement.

Fenêtres temporelles

Toutes les mesures de niveau d'intensité font intervenir la valeur efficace du signal. Or, nous n'avons pas encore abordé la question de la durée T sur laquelle cette valeur efficace est calculée. En théorie, il faudrait tenir compte de toute la durée du signal. En pratique, on a besoin de mesurer une intensité segmentale, pour ne pas attendre un temps infini et pour pouvoir observer l'évolution du niveau d'intensité. Pour avoir une mesure stable sur un signal périodique (Figure 137), il faut mesurer la valeur efficace sur plusieurs périodes. Pour des mesures réglementaires (dB(SPL), dB(A/B/C) ou LUFS), la fenêtre T utilisée est donc en général de quelques centaines de millisecondes (mode *FAST* ou *momentary*), ou de quelques secondes (mode *SLOW* ou *short-term*).

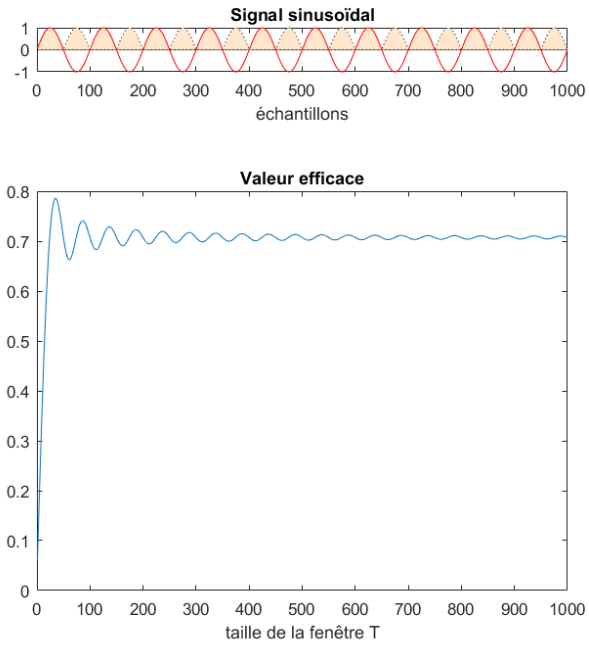


Figure 137 : Evolution de la valeur efficace d'un signal sinusoïdal en fonction de la taille de la fenêtre considérée : on constate que la mesure se stabilise au bout d'une dizaine de périodes (la valeur efficace en T est proportionnelle à la moyenne de l'aire en rose sur [0, T])

Si on s'intéresse à de la parole, des fenêtres plus courtes peuvent être utilisées pour pouvoir visualiser l'enveloppe temporelle du signal. Le logiciel Praat utilise ainsi une fenêtre temporelle de taille variable, T = 3.2 / pitch_minimum, soit quelques dizaines de millisecondes. En revanche, pour estimer l'intensité moyenne d'un locuteur, il faut utiliser des fenêtres temporelles beaucoup plus longues. En effet, la parole est un signal extrêmement variable, donc la mesure d'intensité ne se stabilise réellement qu'au bout d'une dizaine de secondes de parole ininterrompue.

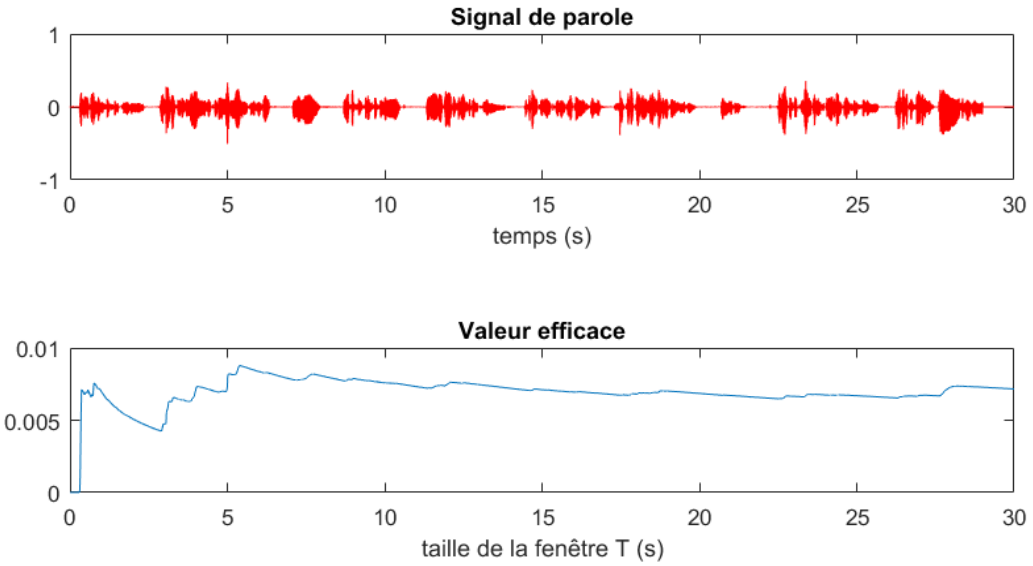


Figure 138 : Evolution de la valeur efficace d'un signal de parole en fonction de la taille de la fenêtre considérée : on constate que la mesure se stabilise au bout d'une vingtaine de secondes

Mesure choisie

Les données recueillies pendant cette thèse étaient numériques. Les mesures effectuées étaient donc nécessairement numériques. Nous avons choisi d'utiliser le dB Praat. En effet, cette mesure semble plus familière, puisqu'elle ressemble à celle utilisée dans les échelles de bruit ; elle est donc plus facile d'accès. Ainsi un son pur d'amplitude maximale a une intensité de 90,97 dB Praat, ce qui en dB A correspondrait à un son pénible, voire dangereux pour des durées d'exposition supérieures à 2h (HCSP, 2013). Un extrait de parole enregistré dans des conditions optimales aura une intensité moyenne d'environ 60 dB Praat, tandis que les silences du même enregistrement auront une intensité moyenne d'environ 20 dB Praat, ce qui correspond aux valeurs attendues respectivement pour une conversation et un environnement très calme.

Par ailleurs, il est intéressant d'utiliser une courbe de pondération, afin de tenir compte des spécificités de l'oreille humaine. En effet, une mesure en dB Praat fournit le même résultat pour un signal sinusoïdal d'intensité fixe quelle que soit sa fréquence : 10 Hz (inaudible), 100 Hz, 1 000 Hz ou 10 000 Hz. Or, nos mesures n'ont pas été faites en chambre sourde, et peuvent être entachées de bruits peu audibles, mais néanmoins présents (le 50 Hz des prises électriques, le bruit de l'autoroute...). Ainsi, la Figure 139 montre un exemple où la simple mesure en dB Praat surestime l'intensité du signal pour certains segments peu audibles.

Nous avons donc choisi d'utiliser la pondération A. En pratique, cette pondération est appliquée avant chaque mesure d'intensité à l'aide d'un filtre numérique implémenté par (Zhivomirov, 2019) d'après les coefficients indiqués dans la norme IEC 61672-1:2002.

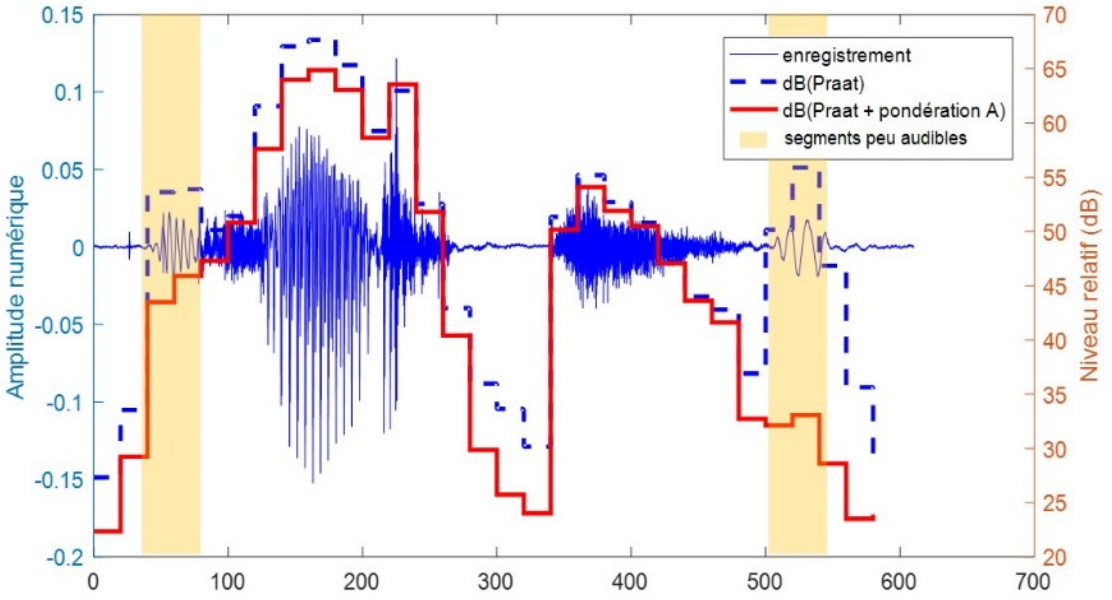


Figure 139 : Comparaison de deux mesures de l'intensité d'un enregistrement de parole
Ici, il s'agit d'une intensité segmentale, mesurée par segments de 20 ms.

Contrairement au logiciel Praat, nous utilisons une taille de fenêtre fixe de 20 ms. Pour obtenir une courbe d'intensité plus régulière, il suffit d'utiliser une fenêtre glissante, plutôt que de mesurer l'intensité de segments disjoints.

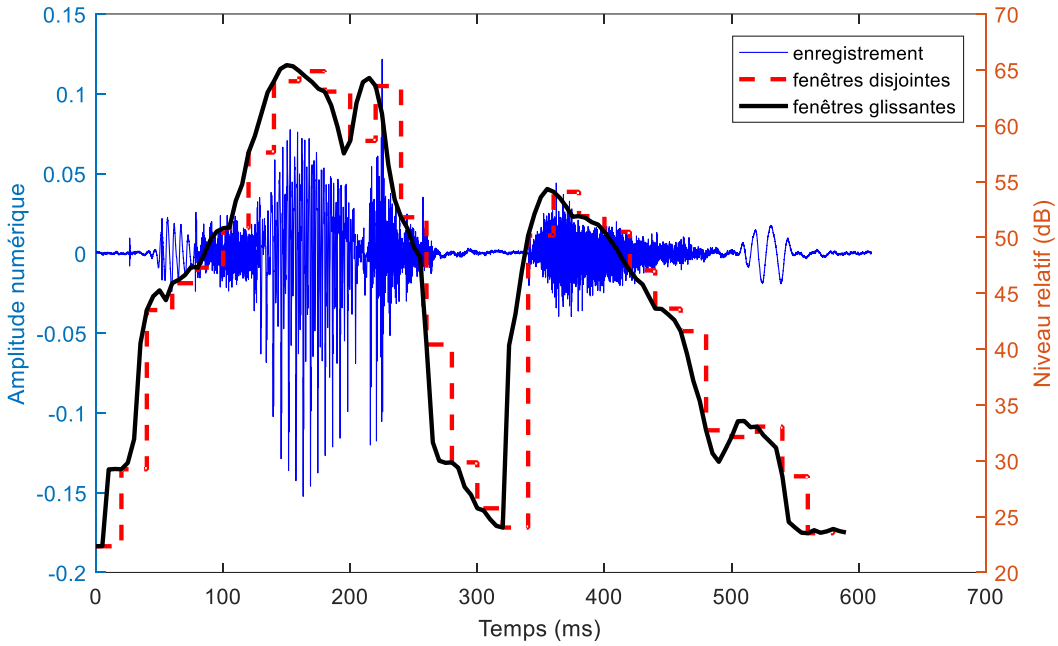


Figure 140 : Comparaison de la courbe d'intensité d'un enregistrement de parole en fonction du type de fenêtre utilisé
Dans un cas, il s'agit de fenêtres disjointes de 20ms. Dans l'autre, il s'agit d'une fenêtre glissante déplacé de 5 ms pour chaque nouvelle mesure.

ANNEXE C :
MODÈLES LINÉAIRES MIXTES

Lorsqu'on analyse des données obtenues dans différentes conditions expérimentales, on cherche généralement à estimer des **effets fixes**, c'est-à-dire l'effet de chaque condition expérimentale sur les données. Les modèles linéaires mixtes permettent de modéliser ces effets fixes, tout en tenant compte d'**effets aléatoires**, pouvant biaiser les résultats de l'analyse. Mathématiquement, ce type de modèle s'écrit comme ceci :

$$y = X\alpha + Z\beta + \epsilon$$

avec :

y :	vecteur des observations
α:	vecteur d'effets fixes que l'on cherche à estimer
β :	vecteur d'effets aléatoires que l'on cherche à estimer
ϵ :	vecteur d'erreurs aléatoires (ce qui reste après modélisation)
X :	matrice de régression contenant les variables des effets fixes
Z :	matrice de régression contenant les variables des effets aléatoires

En R, un tel modèle sera décrit par la formule suivante :

$$\text{grandeur\_mesurée} \sim \text{effet\_fixe} + (1|\text{effet\_aléatoire})$$

Ces modèles sont particulièrement utiles dans le cas où les échantillons étudiés ne sont pas indépendants les uns des autres, ce qui est une des hypothèses fondamentale de l'analyse de variance (aussi appelée Anova). En particulier, dans le cas où plusieurs données appartiennent à un même sujet, ces données ne sont pas indépendantes. Si chaque sujet ne fournit pas exactement le même nombre de données dans chaque condition, il peut apparaître un biais statistique, dont il faut pouvoir tenir compte dans les analyses en considérant que la variable « sujet » est un effet aléatoire.

Pour illustrer cette problématique, nous avons écrit un programme permettant de créer des jeux de données mettant en échec l'analyse de variance (cf. Code 1).

Il s'agit d'une simulation de mesures effectuées sur un groupe constitué initialement de 5 sujets dans deux conditions A et B. Les données ont été tirées aléatoirement à partir d'un générateur de distribution normale, défini par sa moyenne μ et sa variance σ^2. Les paramètres du générateur sont ($\mu = 0$, $\sigma^2 = 20$) en condition A, et ($\mu = 10$, $\sigma^2 = 20$) en condition B. En outre, chaque sujet possède sa propre valeur moyenne (intercept), ajoutée à la donnée fournie par le générateur aléatoire. Le nombre de données tirées est également aléatoire : entre 10 et 20 pour chaque sujet, et chaque condition. Un sixième sujet a ensuite été ajouté de manière à déséquilibrer le jeu de données : sa valeur moyenne est très basse comparée aux autres sujets ; de plus, il a beaucoup plus de données en condition B qu'en condition A.

Ces données sont représentées en Figure 141 (cf. Code 2). Par construction, on devrait pouvoir observer un écart de 10 entre la condition A et la condition B : c'est l'effet fixe, que l'on cherche à modéliser. Or, si on considère uniquement les données globales, cet écart n'est pas visible : au contraire, on observe même la tendance inverse, avec une moyenne plus basse en condition B qu'en condition A. Une simple analyse de variance ne permet donc pas de mettre en évidence l'effet fixe sous-jacent. En revanche, un modèle linéaire mixte permet de calculer non pas la moyenne dans chaque condition, mais les pentes individuelles de chaque sujet entre chaque condition.

Le Code 3 fournit un exemple d'utilisation des modèles linéaires mixtes pour analyser notre jeu de données. Les coefficients du modèle se présentent sous la forme suivante :

	Valeur estimée	Erreur-type
(Intercept)	69,8	18,9
condition B	+ 12,6	3,1

Par défaut, les catégories du modèle sont classées dans l'ordre alphabétique. L'intercept correspond donc à l'estimation de la valeur moyenne en condition A, qui sert de référence. La seconde ligne du tableau correspond à une estimation de la pente entre la condition B et la condition de référence. La dernière colonne correspond à l'erreur-type de ces estimations.

Il faut bien distinguer l'erreur-type mesurée pour l'intercept, et l'erreur-type mesurée pour la condition B. Dans un cas, c'est la valeur moyenne en condition A, qui est connue à environ 19 unités près. Dans l'autre, c'est la pente de la condition B, qui est connue à environ 3 unités près. Ainsi, il ne faut pas interpréter ce modèle en considérant que la valeur moyenne en condition A est de 69,8 ± 18,9 et celle en condition B de 82,4 ± 3,1. Avec une telle incertitude sur la moyenne en condition A, il serait impossible de conclure en l'existence d'un effet fixe. L'erreur-type en condition B est bien une erreur type sur la pente : peu importe que l'écart-type en condition A soit élevé, la pente entre les conditions A et B est estimée à environ 12,6 ± 3,1 : il y a donc bien une augmentation significative entre les deux conditions.

Les modèles linéaires mixtes fournissent également une estimation de la variance expliquée par les effets aléatoires :

	Variance estimée	Erreur-type
sujet	2108,0	45,9
Résidu	384,3	19,6

En première ligne, on trouve une estimation de la variance expliquée par la variable « sujet ». La variance résiduelle apparaît en deuxième ligne : elle représente la part de variabilité qui n'est pas expliquée par le modèle. Ici, elle ne représente qu'environ 15% de la somme des variances ; le modèle est donc bien adapté au jeu de données étudié. Dans le cas d'une modélisation faisant intervenir plusieurs effets aléatoires indépendants, il est possible d'estimer l'importance de chaque effet, en fonction de leur variance respective.

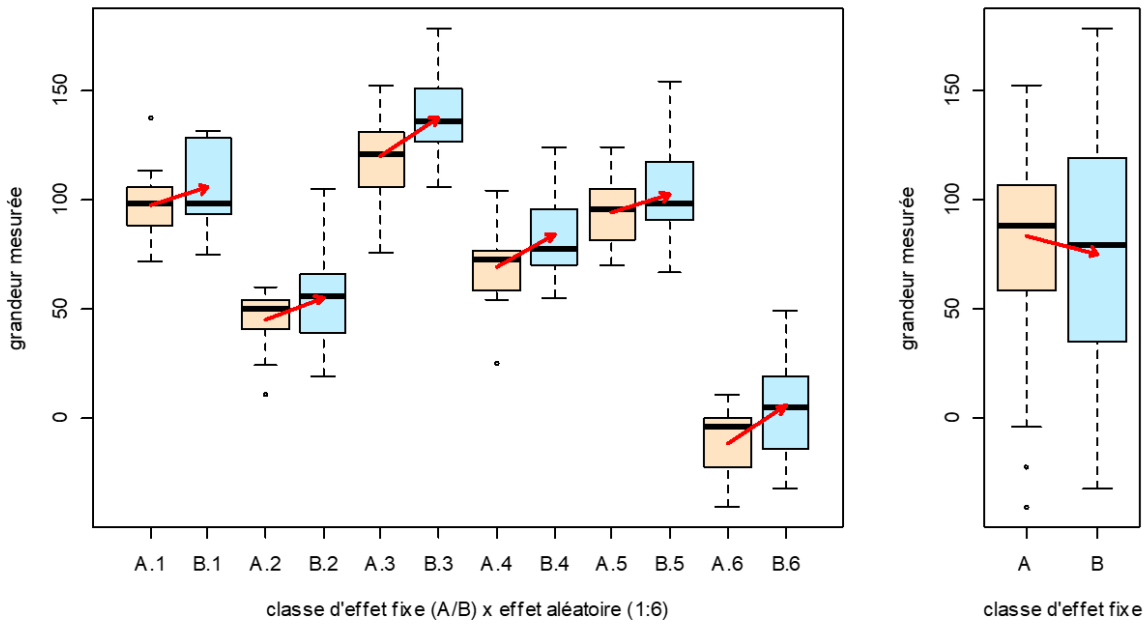


Figure 141 : Exemple de jeu de données (caricatural) où une simple mesure de moyenne est problématique. A et B représentent deux conditions expérimentales, et les chiffres 1 à 6 plusieurs sujets. Les flèches rouges relient les moyennes de chaque groupe.

En observant les résultats sujet par sujet, on constate que la grandeur mesurée est plus élevée en condition B qu'en condition A. Cependant, cette variation n'est pas observable à partir d'un simple calcul de moyenne, car le nombre de données pour chaque sujet n'est pas équilibré. Ainsi, les données du sujet 6 sont beaucoup plus nombreuses en condition B qu'en condition A. Comme la grandeur mesurée est plus faible pour ce sujet que pour les autres, le calcul des moyennes est biaisé : on obtient une moyenne légèrement plus faible en condition B qu'en condition A.

Enfin, pour calculer une valeur-p, on peut faire une analyse de variance pour comparer le modèle étudié (1) au modèle nul (0), qui tient uniquement des effets aléatoires :

$$\text{mesure} \sim 1 + (1|\text{sujet}) \quad (0)$$

$$\text{mesure} \sim \text{condition} + (1|\text{sujet}) \quad (1)$$

Dans notre cas, la valeur-p obtenue est d'environ 10^{-5} : il y a donc seulement 1 chance sur 100 000 que le modèle (0) explique mieux les données que le modèle (1).

Ici, on a considéré que seules les moyennes variaient d'un sujet à l'autre, mais que la pente entre chaque condition était constante. Il est également possible de calculer une pente différente pour chaque sujet, à l'aide du modèle suivant :

$$\text{mesure} \sim \text{condition} + (\text{condition}|\text{sujet}) \quad (2)$$

Code 1 : Création d'un jeu de données problématique pour l'analyse de variance

```
#------------------------------
#--- Création du jeu de données
sujets = 1:5 # numéro des sujets
I = c(100,50,120,80,90) # intercept des sujets
conditions = c('A','B') # conditions d'effets fixes étudiées

df = NULL
for (s in sujets){
  for(c in conditions){
    N = sample(10:20, 1) # nombre de données à tirer pour ce sujet et cette
condition
    for (n in 1:N){
      if (c == 'A'){
        x = rnorm(1, mean=0, sd=20) + I[s] # tirage pour la condition A
      } else {
        x = rnorm(1, mean=10, sd=20) + I[s] # tirage pour la condition B
      }
      df = rbind(df, c(s, c, x))  # ajout des données au tableau
    }
  }
}
colnames(df) = c('sujet', 'condition', 'mesure')

#--- Ajout d'un sujet déséquilibré
# Données du groupe A
for (n in 1:5){
  x = rnorm(1, mean=0, sd=20)
  df = rbind(df, c(6, 'A', x))  # ajout des données au tableau
}
# Données du groupe B
for (n in 1:25){
  x = rnorm(1, mean=10, sd=20)
  df = rbind(df, c(6, 'B', x))  # ajout des données au tableau
}

#--- Finalisation du dataframe
colnames(df) = c('sujet', 'condition', 'mesure')
df = data.frame(df)
df$mesure = as.numeric(paste(df$mesure))
```

Code 2 : Affichage des données

```
#---------------------------------
#--- Affichage
x11(width=5, height=3, pointsize=12)
layout(matrix(c(1,1,1,2), 1, 4, byrow = TRUE))

#--- Détail des résultats sujet par sujet
boxplot(mesure ~ condition*sujet, data = df,
        xlab=" classe d'effet fixe (A/B) x effet aléatoire (1:6)", ylab="grandeur
mesurée",
        col=c('bisque','lightblue1'))

# Calcul des moyennes pour chaque sujet et dans chaque condition
pente = rep(0,6)
A = rep(0,6)
B = rep(0,6)
for (s in 1:6){
  # Sélection des données
  dataA = df[df$sujet == unique(df$sujet)[s] & df$condition == 'A', ]
  dataB = df[df$sujet == unique(df$sujet)[s] & df$condition == 'B', ]
  # Sauvegarde des moyennes
  A = mean(dataA$mesure)
  B = mean(dataB$mesure)
  # Tracé des flèches
  arrows(2*s-1, A, 2*s, B,
         length=0.05, col="red", lwd=2)
}

#--- Résultats globaux
boxplot(mesure ~condition, data = df,
        xlab="classe d'effet fixe", ylab="grandeur mesurée",
        col=c('bisque','lightblue1'))
# Calcul des moyennes globales
dataA = df[df$condition == 'A', ]
dataB = df[df$condition == 'B', ]

# Tracé des flèches
arrows(1, mean(dataA$mesure), 2, mean(dataB$mesure),
       length=0.05, col="red", lwd=2)
```

Code 3 : Modélisation des données à l'aide d'un modèle linéaire aux effets mixtes et par analyse de variance

```
#---------------------------------
#--- Modèles linéaires mixtes
library(lme4)

# Modèle 0 : modèle nul
model0 = lmer(mesure ~ 1 + (1|sujet), data=df)
summary(model0)

# Modèle 1 : le paramètre étudié dépend de la condition
# effets aléatoires pris en compte : sujet
model1 = lmer(mesure ~ condition + (1|sujet), data=df)
summary(model1)

# Comparaison modèles 0 et 1
anova(model0, model1)

#---------------------------------
#--- Anova
fit <- aov(mesure ~ condition, data=df)
summary(fit)
```

ANNEXE D :
LISTE DES QUESTIONS ET CONSIGNES UTILISÉES POUR L'EXPÉRIENCE SUR L'EFFET LOMBARD

Combien y a-t-il de jours dans une semaine ?
De quelle couleur est la violette ?
Combien de pattes a une chèvre ?
De quelle couleur est la planète Mars ?
De quelle couleur est la craie ?
Combien font 2 + 2 ?
Combien de pattes a un cygne ?
Combien font 0 x 7 ?
De quelle couleur est l'or ?
De quelle couleur est la menthe ?
Changer l'apparence de la vidéo
Combien de pattes a un tigre ?
De quelle couleur est la souris ?
De quelle couleur est la citrouille ?
De quelle couleur est le dauphin ?
Combien font 3 + 2 ?
De quelle couleur est la robe d'une mariée ?
Combien de pattes a un poisson ?
De quelle couleur est la cerise ?
Combien de pattes a un moustique ?
Klaxonner
De quelle couleur est le ciel ?
Combien font 1 + 1 ?
De quelle couleur est une émeraude ?
Combien de pattes a une abeille ?
De quelle couleur est la baleine ?
Combien de pattes a une coccinelle ?
Combien font 1 – 1 ?
De quelle couleur est le rhinocéros ?
Combien de pattes a une limace ?
De quelle couleur est le lion ?
Changer l'apparence de la vidéo
Combien font 3 + 3 ?
De quelle couleur est la pistache ?
De quelle couleur est le vin ?
Combien de pattes a un chien ?
De quelle couleur sont les nuages ?
Combien de pattes a un cochon ?
De quelle couleur est la moutarde ?
De quelle couleur sont les piments ?
De quelle couleur est la courgette ?
Combien font 2 x 1 ?
Changer de caméra
De quelle couleur est la colombe ?
De quelle couleur est le rat ?
De quelle couleur est le rubis ?

Combien font 3 x 2 ?
De quelle couleur est le saphir ?
De quelle couleur est le renard ?
Combien de pattes a un pigeon ?
Combien de pattes a un loup ?
De quelle couleur est la banane ?
De quelle couleur sont les brocolis ?
Klaxonner
De quelle couleur est la fraise ?
Combien de pattes a une mygale ?
De quelle couleur sont les haricots ?
Combien font 3 + 5 ?
De quelle couleur est l'éléphant ?
De quelle couleur est le corbeau ?
Combien de pattes a un chat ?
Combien font 3 - 1 ?
De quelle couleur est une voiture de pompier ?
Combien de pattes a un pingouin ?
Quelle est le niveau de la batterie ?
De quelle couleur est la crevette ?
De quelle couleur est l'âne ?
De quelle couleur est la neige ?
Combien de pattes a un cheval ?
De quelle couleur est l'ours polaire ?
De quelle couleur sont les petits pois ?
De quelle couleur sont les ambulances ?
De quelle couleur est l'orange ?
Combien de pattes a une fourmi ?
De quelle couleur est la carotte ?
De quelle couleur sont les yeux du robot ?
De quelle couleur est la mouche ?
Combien font 4 - 4 ?
De quelle couleur est le raisin ?
De quelle couleur est la salade ?
Combien de pattes a une souris ?
Combien font 4 - 1 ?
Combien de pattes a un serpent ?
De quelle couleur est le sanglier ?
De quelle couleur est la tomate ?
Combien de pattes a un âne ?
Changer l'apparence de la vidéo
Combien y a-t-il de jours dans une année ?
Combien de pattes a une tarentule ?
De quelle couleur est une pharmacie ?
Combien font 4 x 2 ?
De quelle couleur est le charbon ?
De quelle couleur est le cochon ?
Combien de pattes a une mouche ?
Combien font 2 + 2 ?
Combien de pattes a un pélican ?
De quelle couleur sont les boutons d'or ?

Est-ce que le réseau a un bon débit ?
De quelle couleur est le sapin ?
Combien de pattes a un escargot ?
De quelle couleur est la mandarine ?
De quelle couleur est le poussin ?
Combien y a-t-il d'œufs dans une boîte de six œufs ?
Combien de pattes a un rat ?
De quelle couleur est l'herbe ?
De quelle couleur est une voiture de police ?
Combien font 6 x 0 ?
De quelle couleur est le castor ?
Klaxonner
De quelle couleur sont les crocodiles ?
Combien de faces a un dé conventionnel ?
De quelle couleur est le tournesol ?
Combien font 5 x 2 ?
Combien de pattes a un poussin ?
De quelle couleur est le loup ?
De quelle couleur est le sel ?
De quelle couleur est la lune ?
De quelles couleurs est un coquelicot ?
De quelle couleur est la planète Terre ?
Changer de caméra
De quelle couleur est la cendre ?
De quelle couleur est la grenouille ?
Combien font 6 – 2 ?
De quelle couleur est un extincteur ?
Combien de pattes a un éléphant ?
De quelle couleur est le papier ?
De quelle couleur est un trombone ?
Combien de pattes a un corbeau ?
De quelle couleur est la rose ?
De quelle couleur est le chou ?
Klaxonner
Combien de pattes a une poule ?
Combien y a-t-il de mois dans une année ?
De quelle couleur est le flamand ?
Combien de pattes a une autruche ?
De quelle couleur est l'abeille ?
De quelle couleur est le citron ?
Combien font 1 + 2 ?
De quelle couleur est le soleil ?
Combien de pattes a une araignée ?
De quelle couleur est la loutre ?
Revenir à l'apparence initiale

RÉSUMÉ

Avec le développement de la robotique grand public apparaît une nouvelle forme de télécommunication : la robotique de téléprésence. Le principe consiste à représenter une personne à distance par l'intermédiaire d'un robot mobile, dont elle peut contrôler librement les déplacements. L'objectif n'est pas simplement de lui permettre de communiquer à distance, mais de lui donner une présence physique et sociale, que le téléphone ou la visioconférence ne suffisent pas à transmettre.

Dans ce contexte, il est particulièrement important de parvenir à transmettre au mieux le « toucher social » du pilote du robot : c'est-à-dire lui permettre d'échanger avec ses interlocuteurs un vaste ensemble de signaux socio-affectifs, qui sont les vecteurs du lien social. En particulier, cette thèse s'intéresse à un élément fondamental du toucher social et fortement impacté par la téléprésence : la portée vocale, à travers laquelle un locuteur contrôle qui peut l'entendre, et s'adapte en permanence aux conditions acoustiques de l'environnement.

À travers une première étude, nous nous intéresserons au lien entre toucher vocal et proxémie, en nous demandant si la manière dont un auditeur perçoit à l'aveugle un interlocuteur dans l'espace peut être influencée par les socio-affects produits par celui-ci. Ensuite, nous montrerons que la portée vocale peut-être affectée par effet Lombard en cas de téléprésence ubiquïte : le pilote, qui perçoit à la fois son environnement local, et l'environnement du robot, s'adapte au niveau de bruit ambiant, même lorsque ce bruit n'est pas perçu par ses interlocuteurs. Enfin, nous présenterons notre participation à un projet Arts et Sciences : le spectacle Aporia, au cours duquel un acteur unique, aidé d'un logiciel de transformation vocale, incarne plusieurs personnages.

ABSTRACT

The development of consumer robotics comes with a new kind of telecommunications systems: telepresence robots. These are mobile robots representing a person who is able to control their movements remotely. The aim is not only to allow remote communication, but to create a sense of social and physical presence, which are not sufficiently transmitted by telephone or videoconferencing.

In this context, it is especially important to ensure that the users' « social touch » is well transmitted, meaning that they are able to exchange a wide range of socio-affective signals, which are the vectors of social links. In particular, this thesis deals with a key element of social touch, which is deeply impacted by telepresence: vocal earshot, by which speakers are normally able to control who can hear them, and to adapt to varying acoustic environment conditions.

In a first study, we will explore the link between vocal touch and proxemics, by asking whether a blind listener's spatial perception of an interlocutor can be influenced by the expressed socio-affects. We will then show that vocal earshot can be modified by the Lombard effect in ubiquitous telepresence, because the pilot is perceiving both the local and remote environments at the same time, and therefore adapts to noise, even if it is not noticeable by the interlocutors. Lastly, we will present our participation in an Arts-Sciences performance called Aporia, during which a unique actor embodies different characters, helped by a voice transforming algorithm.