THÈSE DE DOCTORAT

UNIVERSITÉ DE LILLE

ÉCOLE DOCTORALE BIOLOGIE SANTÉ DE LILLE

L'OBJET « FILTRE VOCAL », DU LABORATOIRE À LA CLINIQUE : VERS L'ANTHROPOTECHNIE DE NOS COGNITIONS SOCIALES

Nadia GUEROUAOU

Thèse co-dirigée par :
Guillaume VAIVA (Lille Neuroscience & Cognition Centre, Lille) &
Jean-Julien AUCOUTURIER (Institut FEMTO-ST, Besançon)

Soutenue le 26/01/24 devant un jury composé de :

Jean-Julien AUCOUTURIER	Institut FEMTO-ST, Besançon	*Co-directeur de thèse*
Anahita BASIRAT	SCALab, Université de Lille	*Examinatrice*
Nicolas BAUMARD	Institut Jean-Nicod, Paris	*Rapporteur*
Baptiste CARAMIAUX	Institut ISIR, Paris	*Rapporteur*
Mathieu TRICLOT	Institut FEMTO-ST, Belfort	*Examinateur*
Guillaume VAIVA	Lille Neuroscience & Cognition Centre	*Directeur de thèse*
Mélanie VOYER	CeRCA, Poitiers	*Invitée*

L'humanité change un peu d'espèce chaque fois qu'elle change à la fois d'outils et d'institutions.

– André Leroi-Gourhan, *Le Geste et la Parole*

Résumé

Entre *zoom calls* et *deep-fakes*, nous vivons aujourd'hui dans un monde marqué par la numérisation croissante de nos interactions sociales, et où nous sommes de plus en plus confrontés à la possibilité de contrôler artificiellement notre apparence visuelle et sonore lors de celles-ci. Cette thèse examine **l'effet de telles technologies de transformation - spécifiquement ici, de « filtres vocaux » capables de contrôler l'expressivité de notre voix - sur les processus cognitifs qui sous-tendent nos perceptions lors d'interactions émotionnelles**.

Nous nous plaçons pour cela dans un double cadre théorique et clinique : d'une part, nous inscrivons notre question au sein de la théorie du traitement prédictif (*predictive processing*), et interrogeons l'effet de contrôler arbitrairement des associations entre états émotionnels et indices expressifs (ex. *je suis heureux, ma voix est souriante*) qui étaient jusqu'alors considérés comme naturels ; d'autre part, nous prenons comme point de départ une situation clinique particulière, la thérapie d'exposition en imagination à l'évènement traumatique chez les patients souffrant de Trouble de Stress Post-Traumatique (TSPT), une situation au contenu émotionnel intense pour laquelle la voix du patient se situe au premier plan.

Ce manuscrit détaille la conception et les résultats de 6 études, qui explorent cette question de façon transverse en invoquant une variété de méthodes expérimentales issues des sciences cognitives. Dans le Chapitre 2, nous présentons une étude d'**éthique expérimentale** interrogeant les facteurs qui influencent l'acceptabilité morale de l'usage de « filtres vocaux » dans différentes situations hypothétiques. Nous y montrons une grande acceptabilité morale des filtres permettant de modifier la tonalité émotionnelle vocale, notamment pour un usage thérapeutique et afin d'atténuer les émotions de valence négatives. Dans le Chapitre 3, nous présentons deux **études cliniques longitudinales** menées chez des patients souffrant de TSPT, et grâce à une méthodologie d'analyse acoustique informatique, mettons en évidence que la voix des patients - et en particulier leur fréquence fondamentale (*pitch*) - reflète non seulement l'abaissement de la symptomatologie au cours de la thérapie mais aussi l'évolution de leur rythme cardiaque. Dans le Chapitre 4, nous présentons une **étude de perception** qui interroge la capacité de soignants, spécialistes ou non du TSPT, d'évaluer la symptomatologie des patients en fonction de leur voix. En utilisant le

« filtre de pitch » comme une méthode de contrôle expérimental des stimuli, nous démontrons une influence causale du pitch sur ces inférences médicales - et, le cas échéant, la capacité d'un tel filtre à tromper ces jugements. Enfin, dans le Chapitre 5, nous présentons deux études visant à interroger causalement le lien entre pitch de la voix et rythme cardiaque (RC) observé de façon corrélationnelle au Chapitre 3. Nous présentons tout d'abord une **étude de psychophysiologie**, dans laquelle nous manipulons causalement le RC de participants avec une intervention inspirée du protocole clinique de *tilt test*, et étudions l'influence de cette manipulation sur le pitch de leur voix enregistré simultanément. Nous présentons ensuite une **étude de perception**, dans laquelle nous interrogeons la capacité d'observateurs à inférer le RC de locuteurs à la simple écoute de leur voix. Nous montrons que, même si la tâche est difficile, les participants arrivent dans les meilleures conditions à associer le bon RC à une voix donnée, avec des performances meilleures qu'au hasard ; et qu'ils utilisent pour cela une heuristique liant pitch et RC, même lorsqu'elle ne correspond pas à la réalité.

Pris dans leur ensemble, les résultats de ces 6 études confirment l'influence du « filtre de pitch vocal » sur les processus d'inférences perceptives qui sous-tendent nos cognitions sociales fines en situation d'interaction. Nous discutons du potentiel anthropotechnique de ce nouvel objet technologique, et de la nécessité de poursuivre l'étude des effets des nouvelles technologies de façonnement du soi sur la cognition, en vue de l'élaboration d'une pensée critique à la hauteur des enjeux philosophiques, scientifiques et technologiques que posent ces nouvelles technologies.

Remerciements

La part des autres,

Les premières pages de mon manuscrit sont dédiées à tous ceux qui ont contribué, de près ou de loin, à la réalisation de ce travail. A ces autres qui nous font progresser en nous éveillant à des questions et ainsi un peu à nous-mêmes.

J'aimerais remercier mes directeurs de thèse, Jean-Julien Aucouturier et Guillauma Vaiva. En premier lieu pour la grande confiance qu'ils ont accordé au sujet de thèse que je leur proposais initialement, ensuite pour la liberté qu'ils m'ont laissée de l'explorer de manière hautement inter-disciplinaire et enfin pour les échanges approfondis, rigoureux et éclairants qui m'ont stimulée pendant ces quatre années, tout particulièrement nos discussions passionnées au sujet de chacun de nos design d'expérience. Merci pour tout cela, et plus encore.

Je tiens également à remercier l'équipe PDS de l'Ircam pour son accueil on ne peut plus chaleureux. De manière synthétique et non exhaustive, Emmanuel Ponsot pour ses précieux conseils, Olivier Houix pour ses recommandations musicales, Patrick Susini son goût du beau, et Nicolas Misdariis pour, entre autres, sa grande générosité et son écoute bienveillante. Cette thèse n'aurait pas été la même sans notre promo de PhD, Constance, Claire, Valérian, Yann, Clément, Paul, Lenny, Victor, Pablo, Vincent [...] et Baptiste (mon presque jumeau de soutenance). Un mot évidemment à mes collègues qui ont richement et agréablement collaboré aux travaux de cette thèse. En recherche : Coralie Vincent, Paul Maublanc et Matthieu Fraticelli. En clinique, les psychologues et psychiatres du CRP des Hauts de France, et tout spécialement à Frédérique Warembourg, notre cheffe de service, pour sa grande sollicitude à mon

égard ainsi qu'à Séverine Vanhoove, Stéphane Duhem, Alice Damarey et Ludivine Gaultier pour leur participation cruciale aux études de cette thèse.

Aussi, je tiens à exprimer ma profonde gratitude au Pr Katz Watanabe et son équipe de m'avoir accueillie dans leur laboratoire à Tokyo pour deux mois inoubliables. Je remercie également tous les chercheurs qui ont pris part à mon travail sur place (Pr Takashi Ikegami, Pr Philippe Codognet, Pr Koichiro Eto, Pr Yuko Yotsumoto, Pr Olaf Witkowsky, Dr Shigeo Yoshida et Elena Knox).

De manière plus personnelle, toute ma gratitude va aussi à Inès Weber, Abdennour Bidar et Caco Olezac pour leurs inestimables conseils dans ce travail de rédaction ainsi que pour avoir accueilli ma réflexion au sujet de l'écriture comme processus créateur et transformateur de soi.

Pour terminer, mes remerciements les plus chaleureux vont à mon père, pour sa bonté et son soutien de toujours et à mes amis (Ama, Zico, Jimmy, Néri, Bart, Clélie, Chouch, Bib, Jules, Julie, Fernanda...) pour leurs témoignages d'affection et d'intérêt pour ma thèse pendant ces années de travail, à Poupie pour avoir adorablement composé son emploi du temps de façon à préserver mon sommeil et ma concentration. Et tout particulièrement à Emilien pour ses encouragements sans faille, qui ont été essentiels pendant ces longs mois et Paùla, pour sa présence indispensable à mes côtés pendant cet été de rédaction qu'elle a rempli de délicates attentions. Un mot pour Gogo, dont la détermination et la passion à l'égard de la vie ont été de véritables sources qui ont accompagné l'écriture de ce manuscrit.

Enfin, je remercie une émission de radio et un livre (à défaut de pouvoir remercier leurs auteurs). *La Conversation Scientifique*, par Etienne Klein, qui a choisi pour programme estival -inspiré de Bachelard- l'exploration de l'imagination, l'imaginaire et le rêve dans l'esprit des scientifiques, dont l'écoute a considérablement aidée mes choix d'écriture et les "*Exercices spirituels et philosophie antique*" de Pierre Hadot, un ouvrage qui m'a fidèlement accompagnée durant ces mois de rédaction et auquel je dois beaucoup.

Table des matières

1. Introduction : de Darwin au Deep-learning

1.1 Cadre général

Les émotions revêtent une fonction particulière pour les êtres humains. Leurs expressions faciales et vocales peuvent être qualifiées selon une perspective psycho évolutionniste de « signaux », c'est à dire des manifestations comportementales ayant pour objet d'informer et parfois manipuler autrui au sujet de nos états internes (Darwin, 1872 ; Knapp et al., 2013). En modulant continuellement l'activité de nos muscles faciaux ainsi que les structures phonatoires et articulatoires de notre appareil vocal, nous fournissons en effet un second canal, non linguistique, à nos conversations quotidiennes. Ce canal véhicule des informations permettant à nos interlocuteurs d'interpréter nos états émotionnels tels que la joie ou la surprise (Jack et al., 2016), nos attitudes sociales telles que la bienveillance ou la domination (Ponsot et al., 2018), ou encore des éléments « métacognitifs », comme la certitude ou le doute (Goupil et al., 2021). À partir de ces signaux paralinguistiques, et en s'appuyant sur d'autres indices comme le contexte de l'interaction en cours, notre interlocuteur va se faire une idée de notre état émotionnel.

Cette théorie récente d'une véritable « construction » cognitive des émotions (Barrett, 2012), est de plus en plus répandue au sein des neurosciences. En effet, si nos expressions faciales et vocales ont été façonnées par une longue et délicate évolution biologique, leur fonction de communication des émotions dépend tout autant de l'histoire culturelle de l'environnement au sein duquel elles sont exprimées (Jack et al., 2012 ; Safra et al., 2020). Ainsi, le fait d'avoir développé notre cognition sociale dans une culture plutôt qu'une autre peut nous amener à prêter attention à un indice facial

ou un autre, voire à interpréter ceux-ci comme des expressions d'émotions différentes (Jack et al., 2012). Similairement, le contexte historique d'une société peut amener celle-ci à biaiser le niveau de base d'attitudes comme la confiance ou la défiance dans ses représentations picturales (Safra et al., 2020), et à l'inverse on peut interroger ce que la diffusion de ces représentations picturales dans la société engendre sur les inférences que font les individus qui la composent.

Dès lors, pour nos émotions comme pour l'ensemble de notre cognition, nous sommes donc les descendants des artefacts culturels et en particulier des objets techniques parmi lesquels nous vivons. Ainsi, selon Nishida Kitaro, *« la technique est l'expression de l'esprit humain interagissant avec son environnement et qui, par cette interaction, se forme lui-même »*. A cela il ajoute que *« nous créons des objets par le biais de la technique et qu'en les créant, nous nous créons nous-mêmes »* (Lennerfors and Murata, 2019). Les avancées technologiques considérables dont l'occident est témoin ces dernières années pourraient alors modifier radicalement la façon dont nous utilisons les signaux paralinguistiques dans nos interactions quotidiennes. Les progrès récents en matière de techniques de traitement du signal ont en effet rendu possible la manipulation en temps réel d'expressions faciales telles que les sourires (Arias et al., 2020) et d'indices vocaux tels que la hauteur (Rachman et al., 2017) ou le timbre (Arias et al., 2020).

De manière peut-être encore plus radicale, le développement dans ces dernières décennies d'architectures de réseaux de neurones profonds dans le domaine de l'apprentissage machine, a ouvert la possibilité de manipuler de manière paramétrique des actions faciales individuelles (Pumarola et al., 2018) et de *convertir une voix en plusieurs variantes émotionnelles* (Luo et al., 2017), ou pour le dire autrement, de créer des deepfakes (ou « hypertrucages »). La part de plus en plus importante des communications en visioconférence depuis la pandémie de Covid-19 participe de l'installation récente de ces formes d'expressions émotionnelles « synthétiques » dans notre quotidien.

Partant de cet état de fait, et sous-tendue par la méthodologie des neurosciences cognitives et de l'éthique expérimentale, cette thèse se propose de réfléchir au devenir de notre capacité de perception des émotions, façonnée par la culture, dans un monde où les objets techniques permettent dorénavant de créer et contrôler artificiellement des indices expressifs faciaux et vocaux jusqu'alors considérés comme « naturels » au

moyen d'objets que nous serons amenés à appeler « filtres » dans le corps de cette thèse.

A cette fin, notre travail prendra pour objet la transformation informatique des émotions dans la voix et en particulier son utilisation dans le cadre du soin en psychiatrie chez les patients souffrant de Trouble de Stress Post Traumatique (TSPT). Parce que ce filtre vocal est un objet technique, son étude selon Simondon (1989) nécessiterait d'*« adopter le point de vue conjoint de l'ingénieur qui invente, du sociologue qui étudie les usages sociaux et du psychologue qui analyse les processus psychiques concomitants de l'inventeur, de l'usager et de la réalité humaine déposée dans le système technique pour parvenir à saisir le sens du fonctionnement et du développement d'une machine »*. Sans prétendre parvenir à regrouper ces différentes disciplines au sein de ce manuscrit, nous emprunterons donc néanmoins certains concepts à la philosophie des techniques, et aborderons les aspects sociétaux intrinsèques au sujet de notre travail. Nous souhaitons ainsi alimenter notre questionnement auquel nous nous emploierons à répondre en nous appuyant sur une méthodologie expérimentale qui correspond au mode des neurosciences cognitives.

1.2 Emergence des deepfakes

Le Centre pour l'Ethique des Données et l'Innovation (CDEI) définit les « deepfakes » comme le résultat d'une *« technique de synthèse d'images basée sur l'intelligence artificielle qui consiste à créer un contenu vidéo falsifié mais très réaliste, grâce à laquelle il est possible de modifier la façon dont une personne, un objet ou un environnement est présentée »* (GOV.UK, 2019). En plus de modifier le contenu visuel, *« les enregistrements audio de la voix d'une personne [...] peuvent être utilisés pour recréer leur discours et leur faire dire n'importe quelle phrase »* (Vaccari and Chadwick, 2020). On distingue ainsi les deepfakes consistant en la génération ou la modification de visages et ceux visant la synthèse vocale, en somme les deepfakes *visuels* et les deepfakes *vocaux*, ces derniers étant l'objet de la présente thèse. Pour autant,les deepfakes visuels connaissant un peu d'avance sur ceux visant uniquement la voix, (ou les deux étant parfois entremêlés quand il s'agit de vidéos) nous développerons dans cette partie des exemples illustrant les deux modalités visuelle et vocale. Le traitement du signal et l'apprentissage machine visuels ayant traditionnellement une longueur d'avance sur l'audio (Vincent,

2019), nous motivons ce choix par l'intuition que certains usages populaires dans le domaine visuel préfigurent ceux dans le domaine vocal et nécessitent donc d'être présentés.

1.2.1 Une diversité d'applications

Face à ces possibilités techniques nouvelles, et au siècle du « numérique intégral » de Stiegler (2016b), un vaste éventail d'usages, plus ou moins souhaitables, se dessine. Les usages « malveillants » associés à cette technique fille de l'IA sont les plus connus du grand public car c'est par leur fait que la technologie a été médiatisée. L'utilisation malveillante de deepfakes est étroitement liée à la manipulation de l'information et aux campagnes de désinformation, ainsi qu'à la diffusion de fausses informations et de théories conspirationnistes (for Parliamentary Research Services., 2019). Parmi les exemples populaires récents, citons celui du deepfake vidéo du président ukrainien Zelensky diffusé au début de la guerre en Ukraine, dans lequel il apparaît exhorter ses troupes à se retirer, ou encore la vidéo manipulée de Nancy Pelosi 1.1, principale opposante à Donald Trump, massivement relayée lors de la campagne présidentielle américaine de 2019 (vidéo consultable via le lien suivant `https://youtu.be/F7_DBM7OIfM`). Celle-ci montre l'élue américaine pendant un discours au Centre Américain pour le Progrès s'exprimer avec difficulté, son débit au ralenti et l'articulation laborieuse.

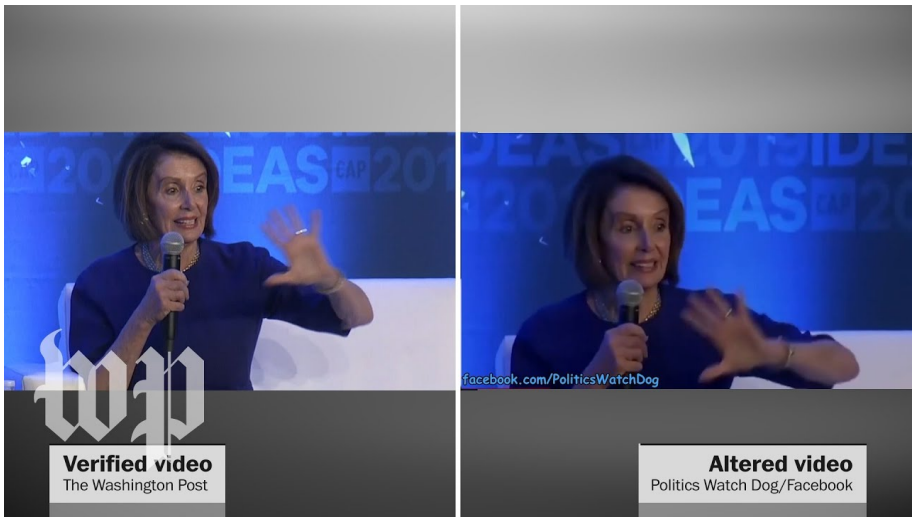


FIGURE 1.1 – Exemple du shallow-fake de Nancy Pelosi

Ce « shallow-fake » (ainsi appelé car la manipulation technologique est relativement grossière et réalisé sans apprentissage machine, par opposition à un fake plus « profond ») a permis en ralentissant le débit de parole de la candidate et en corrigeant la hauteur de sa voix de donner l'impression qu'elle était ivre. Le fait que ce seul procédé technique, visant uniquement la voix de la femme politique, ait eu un tel impact sur la réception de son discours, alors même que son contenu verbal reste inchangé, témoigne de l'importance des éléments paralinguistiques dans la communication.

Cependant, l'utilisation de deepfakes peut également être motivée par des objectifs satiriques ou à des fins culturelles et de divertissement. A l'instar d'Alan Zucconi, maître de conférence en intelligence artificielle à l'Université Goldsmiths de Londres qui enseigne ce qu'il nomme les « applications positives » potentielles des deepfakes, comme les reconstitutions historiques, le doublage plus réaliste des films en langue étrangère, la recréation numérique du membre d'un amputé ou la possibilité pour les personnes transgenres de se projeter en image dans autre sexe. En France, dans le domaine culturel, la technique de clonage vocal a déjà été mise en œuvre pour recréer les voix de deux personnalités défuntes en collaboration avec des média bénéficiant d'une très large audience : la voix de Dalida utilisée pour le programme télévisé « Hôtel du temps » de Thierry Ardisson diffusé sur France 3 et celle du Général de Gaulle afin de reconstituter le discours du 18 juin pour Le Monde. Ce genre de prouesse repose sur diverses techniques de traitement de la voix. Cependant, afin de mettre en exergue leur proximité avec les objets visuels qui sont déjà largement déployés et connus dans notre société, nous utiliserons dans ce travail le terme de « filtre » pour renvoyer à ces transformations de voix réalistes ou deepfake vocaux même si leur implémentation ne relève pas forcément du concept de « filtrage » en traitement du signal.

En électronique, un filtre numérique est un élément qui modifie le contenu spectral d'un signal d'entrée en amplifiant ou atténuant certaines composantes spectrales (https://fr.wikipedia.org/wiki/Filtre_(%C3%A9lectronique). Si certaines manipulations de voix peuvent être réalisées en utilisant de tel filtres (par ex., le travail de Rachman et al. (2017) montre qu'un filtre « low-shelf » qui amplifie les basses fréquences peut rendre une voix plus triste ou moins énergétique), la plupart des manipulations considérées dans cette thèse ne relèvent pas stricto-sensu d'une technique de filtrage. C'est notamment le cas de la transformation de hauteur (pitch en anglais) utilisée tout au long de notre travail, qui utilise un algorithme dit de « vocodeur de phase » qui analyse le signal avant de le modifier à la fois dans sa dimension spectrale et temporelle. C'est également le cas *a fortiori* des nouvelles méthodes issues de l'apprentissage profond, dont le traitement très fortement non-linéaire n'a vraisemblablement pas de représentation équivalente sous la forme d'un filtre, aussi complexe soit-il. Cependant, nous utiliserons dans tout le reste de ce travail le terme « filtre » pour renvoyer à de telles manipulations vocales, à la fois par convenance d'écriture, par analogie avec le terme consacré (et tout aussi impropre) dans le domaine des filtres visuels d'outils comme Zoom, et pour noter que conceptuellement ces outils permettent malgré tout d'*amplifier* ou *atténuer* certaines composantes expressives ou émotionnelles de la voix (ex. filtre de sourire, filtre de pitch), même si celles-ci ne se réduisent pas à de simples fréquences localisées dans le spectre.

1.2.2 Du divertissement au façonnement de soi : l'exemple du filtre à selfie

A l'heure où nous écrivons cette thèse, si les précédentes utilisations des techniques d'apprentissage profond bénéficient d'une large exposition notamment via les réseaux sociaux, c'est sans doute leur expression sous la forme des filtres à selfies qui connaît le déploiement le plus impressionnant. Les plateformes de médias sociaux visuels, comme Instagram (Facebook Inc., California, USA) et Snapchat (Snap Inc, Santa Monica, CA)), deux plateformes de partage de contenus photo et vidéo, dont la popularité chez les 15-25 ans n'a cessé de croître au cours des quinze dernières

années ont ainsi introduit des filtres de réalité augmentée (RA) en 2015 qui sont alors devenus une fonctionnalité très répandue pour prendre des photos de soi (*selfies* en anglais). Six-cent millions de personnes les utilisent ainsi chaque mois sur Instagram ou Facebook et 76 % des utilisateurs de Snapchat les appliquent chaque jour (Javornik et al., 2022). Certains de ces filtres relèvent du ridicule, comme des oreilles de chat ou des personnages fantastiques, tandis que d'autres permettent aux utilisateurs de façonner numériquement leur visage pour se conformer à des normes de beauté spécifiques. Ces filtres de beauté en réalité augmentée ne se contentent pas d'appliquer un maquillage numérique, ils vont plus loin en déformant la mâchoire et le nez de l'utilisateur, en agrandissant ses yeux et ses lèvres et en lissant sa peau 1.2. Avant le développement des filtres, les selfies ne pouvaient être améliorés que par une retouche photo rétroactive. Aujourd'hui, la modification du visage est possible grâce à des filtres qui s'adaptent aux caractéristiques du visage en temps réel.

.

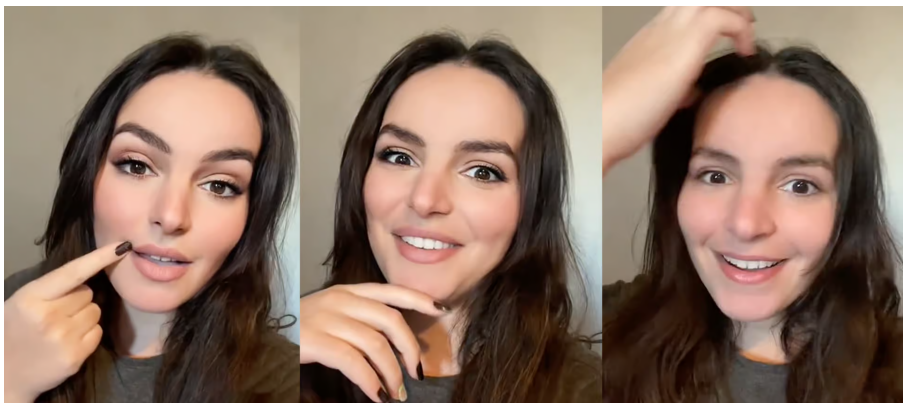


FIGURE 1.2 – La TikTokeuse `@sabrina__chan` fait une démonstration du filtre « Bold Glamour » sur TikTok. (https://www.futura-sciences.com/tech/actualites/intelligence-artificielle-bold-glamour-ce-filtre-hallucinant-tiktok-fait-polemique-104072/)

Une étude menée au Royaume-Uni en 2021 par Rosalind Gill (Orgad and Gill, 2021) professeure d'analyse sociale et culturelle, a révélé que 90 % des jeunes femmes

appliquent des filtres de beauté ou modifient leurs photos avant de les publier sur les médias sociaux. Ce chiffre est à prendre d'autant plus au sérieux que l'usage de ces filtres semble impacter les comportements et états mentaux des utilisateurs au delà même de l'écosystème numérique et créent chez certains une confusion quant à leur corps réel. Plusieurs phénomènes sont ainsi documentés à cet égard, comme l'apparition de la « dysmorphie du selfie » définie par l'apparition chez les habitués des selfies de problèmes d'estime de soi et de distorsion corporelle (Javornik et al., 2022), exacerbés par l'utilisation de filtres jusqu'à mener dans certains cas extrêmes au recours à la chirurgie esthétique. *« Les chirurgiens plasticiens ont constaté une augmentation constante de l'intérêt pour l'augmentation invasive et non invasive du visage, en particulier chez les jeunes patients, affichant souvent des attentes irréalistes en raison des ajustements d'un filtre »* (Daar et al., 2021). De plus, l'enquête annuelle 2018 de l'American Academy of Facial Plastic and Reconstructive Surgery confirme le rôle croissant des selfies dans la chirurgie plastique du visage, en particulier dans les procédures non chirurgicales, le désir d'apparaître plus beau dans un selfie étant une motivation principale. Cette tendance s'est surtout manifestée chez les Millennials, âgés de 22 à 37 ans et est loin d'être triviale (Cristel et al., 2021). Il convient de relever qu'ici, le souhait des patients de procéder à la chirurgie plastique est motivé par un désir d'améliorer leur image numérique. Nous serions en quelque sorte alors témoins de la survenance d'une forme d'inversion des pratiques : l'apparence réelle en vient à être modifiée pour améliorer celle d'un soi numérique. En résumé, il semblerait donc que nous soyons donc face à un double phénomène : le filtre, de par l'image améliorée qu'il renvoie crée d'une part un sentiment d'insatisfaction quant à son apparence réelle, menant au souhait de chirurgie pour se conformer à cette apparence idéalisée créée algorithmiquement et d'autre part, des transformations physiques chirurgicales sont sollicitées afin cette fois, d'« améliorer » non plus son apparence physique réelle mais l'image numérique, alors même que cette dernière ne reflète pas la réalité car déformée par l'objectif de l'appareil photo. Il a en effet été établi qu'un selfie augmente la largeur du nez de 30 % (Ward et al., 2018), une donnée méconnue du grand public alors que l'apparence du nez est une plainte fréquente chez les patients lorsqu'il est montré sur des selfies en consultation de chirurgie plastique.

De façon symétrique, l'utilisation de filtres visuels sur les réseaux sociaux peut aussi avoir un effet positif sur le bien-être psychologique des usagers, par exemple en

mettant en avant des aspects de soi qui sont moins visibles au quotidien ou en permettant l'exploration de son identité. Selon Javornik et al. (2022) l'impact psychologique de ces objets sur la santé mentale et le bien-être dépend des motivations sous-jacentes à leur utilisation : « les personnes qui utilisent les filtres à des fins de divertissement, d'interaction sociale ou pour transformer leur soi numérique bénéficient souvent de cette technologie, et un effet positif sur l'humeur peut être observé. Cependant, l'utilisation de filtres de beauté pour idéaliser le soi et fausser son image conduit souvent à une moins bonne acceptation de soi et a des effets négatifs sur la perception que les utilisateurs ont d'eux-mêmes ».

Les réseaux sociaux ne sont pas les seuls à proposer l'usage de ce type de filtres visuels. Il en est de même pour la plateforme de vidéo conférence Zoom, qui propose sous l'onglet `améliorer la mauvaise qualité vidéo`, l'utilisation de filtres personnalisés afin d'*« améliorer l'expérience des réunions »* ou encore dans la rubrique `Sentez-vous plus en confiance en vidéo` où l'on peut lire « Préparez-vous encore mieux à la vidéo grâce au contrôle granulaire de l'intensité de vos retouches et ajustements lumineux, pour un bon éclairage quelle que soit l'exposition. Changez la luminosité de votre écran et le degré de lissage de la peau pour optimiser votre apparence en vidéo ! » proposée sur la page officielle du blog de Zoom `https://blog.zoom.us/fr/filters-reactions-lighting-features-zoom-meetings-2/`. Par ailleurs, il faut noter concernant cette plateforme que son utilisation semble également susciter des troubles de l'image de soi en lien avec l'effet miroir implémenté dans l'outil. La « dysmorphie Zoom » a notamment récemment vu le jour et, est elle aussi associée à une recrudescence de demandes de chirurgie plastique (Daar et al., 2021). A ce sujet, les chirurgiens esthétiques déclarent que leurs patients travaillant principalement à domicile par vidéoconférence ont confié prêter plus attention aux traits de leur visage qu'ils trouvent les moins attrayants et être de plus en plus gênés par ces « zones à problèmes » particulières. Ces professionnels voient en cela un réel sujet à adresser spécifiquement : *« Nous devons tenir compte de ce nouveau contexte de quarantaine, du manque d'interaction sociale en personne et de la montée en puissance de la culture Zoom pour adapter la manière dont nous évaluons et soignons nos patients ayant recours à la chirurgie esthétique »* (Daar et al., 2021).

En plus de permettre différentes retouches quant au traits physiques des visages, les filtres permettent également d'afficher une expression faciale émotionnelle simulée à l'instar du sourire artificiel illustré en Fig. 1.3.

FIGURE 1.3 – Illustration de l'effet « smile » d'édition de selfie proposé par l'entreprise *Perfect Corp* sur leur site internet le 20 juillet 2023.

Il convient alors de souligner que le « smile design » qui renvoie à la définition et au façonnement d'un sourire considéré comme beau est en partie modelé par les normes culturelles (Davis, 2007). Or, les systèmes d'IA générative à l'origine de ces productions artificielles étant dominés par des sources d'images standardisées sur le modèle du sourire suivant l'esthétique nord-américaine (voir Fig.1.4), nous observerions, sur un plan anthropologique, l'émergence d'une nouvelle monoculture visuelle des expressions faciales (voir Fig.1.5) suscitant l'inquiétude de certains chercheurs qui prédisent notamment que la diversité de l'expression humaine ne survivra pas à l'hégémonie des algorithmes. Ainsi selon Gurfinkel (2023), *« en aplatissant la diversité des expressions faciales des civilisations du monde entier, l'IA a réduit le spectre de l'histoire, de la culture, de la photographie et des concepts d'émotion à une perspective singulière et monolithique. Elle a présenté un faux récit visuel sur l'universalité de quelque chose qui, dans le monde réel - où de vrais humains ont vécu et créé une culture, une expression et une sémiologie pendant des centaines de milliers d'années - est tout sauf uniforme. »*.

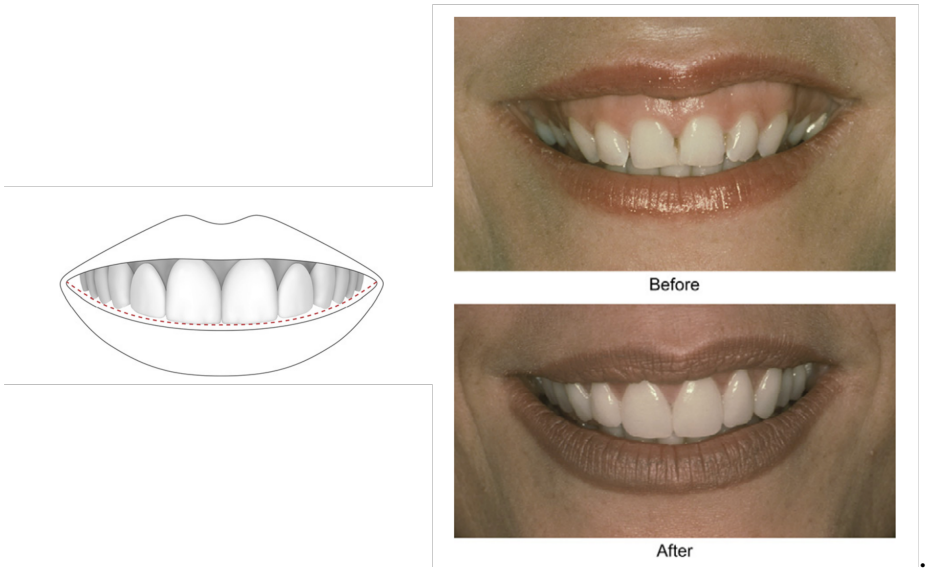


FIGURE 1.4 – Figures tirées de l'article de Davis (2007). A gauche : « Ligne du sourire idéal. (Avec l'autorisation de l'American Academy of Cosmetic Dentistry, Madison, WI) ». A droite : Procédures dentaires cosmétiques visant à designer un sourire « plus esthétique » : « Avant (en haut) et après (en bas) l'allongement de la couronne, la frénectomie et la pose de 10 facettes en porcelaine au maxillaire »

FIGURE 1.5 – Photographie générée par l'IA générative midjourney sous le prompt « samourai warrior selfie » tiré de l'article de Gurfinkel (2023)

Alors que, comme nous l'avons vu, notre image se façonnerait de plus en plus dans les reflets de l'IA, encore peu d'études à notre connaissance ne semblent adresser directement les effets que ces sourires ou autres expressions artificielles pourraient engendrer sur les processus cognitifs visés par ces transformations, à savoir la perception des émotions chez les individus qui y sont exposés. Dans cette veine, bien que ne traitant pas la question des émotions, on peut néanmoins citer les travaux de Nightingale and Farid (2022) ; Tucciarelli et al. (2022), menés sur des photographies de visage entièrement générées par une IA générative StyleGAN2 (Karras et al., 2020), qui mettent en évidence que ces visages artificiels sont considérés comme plus dignes de confiance (voir Fig. 1.6, Nightingale and Farid (2022)) et plus réalistes (Tucciarelli et al., 2022) que des photos d'individus réels.

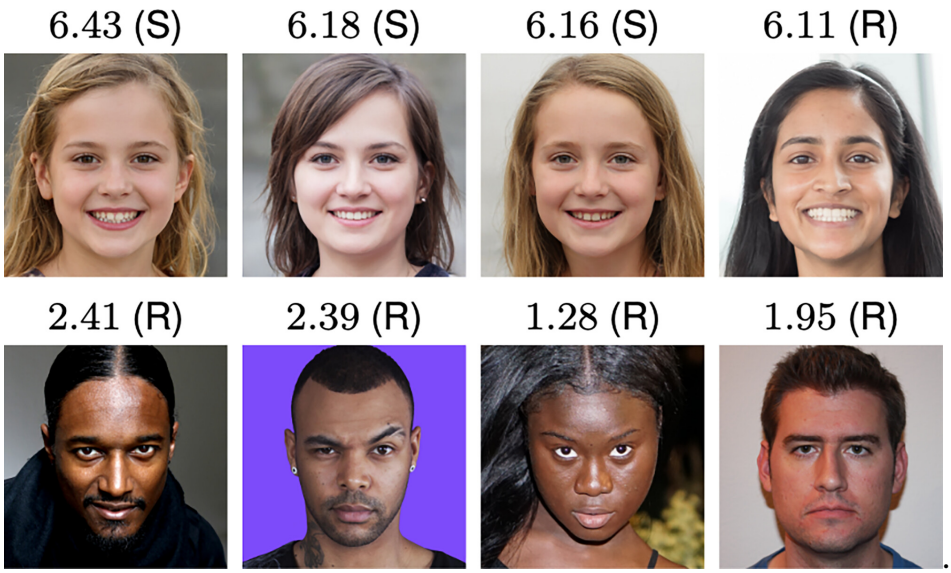


FIGURE 1.6 – Figure tirée de l'article de Nightingale and Farid (2022). Les quatre visages les plus dignes de confiance (en haut) et les quatre visages les moins dignes de confiance (en bas) et leur note de confiance sur une échelle de 1 (très peu digne de confiance) à 7 (très digne de confiance). Outre un évident biais de genre et d'ethnicité sur cet exemple, les visages synthétiques (S) sont, en moyenne, jugés plus dignes de confiance que les visages réels (R).

1.2.3 Le filtre vocal, nouvelle technologie du soi ?

Si à la différence des filtres à selfie que nous venons de décrire, les filtres vocaux ne font pas encore partie intégrante de notre écosystème numérique, la popularité de leurs cousins visuels nous invite à considérer avec attention cet objet dès maintenant afin d'anticiper leur déploiement futur potentiel.

Si l'on s'intéresse aux travaux en laboratoire, incubateurs des technologies d'augmentation de demain, différentes techniques informatiques ont vu le jour permettant aujourd'hui de changer l'identité d'une voix en modifiant par exemple sa taille, son sexe et son âge apparents (Farner et al., 2009), ou des aspects plus qualitatifs en la rendant plus douce ou plus rauque (Degottex et al., 2013,1). De manière plus dynamique, il est également possible de convertir un extrait vocal en plusieurs variantes exprimant d'autres émotions (Beller, 2010) ou des attitudes sociales diverses (Moine and Obin, 2020). Ces techniques vont jusqu'à la possibilité de « clonage vocale », lorsque la voix cible souhaitée est spécifique à une personne (Abe et al., 1990 ; Toda et al., 2007).

Le panorama de ces différentes techniques évolue très rapidement (voir Mohammadi and Kain (2017) ; Stylianou (2009) pour une *review*). Pour n'illustrer par exemple que les transformations visant l'expression de l'émotion en temps réel, et dont on trouvera une review récente dans Arias et al. (2021), un dispositif conçu au sein de l'équipe de recherche dans laquelle s'inscrit cette thèse permet de modifier en direct la tonalité émotionnelle du discours formulé à l'oral : le DAVID (abréviation pour 'Da amazing voice inflection device'). Il s'agit d'une plateforme disponible en téléchargement open-source dans l'environnement `Max` (©Cycling '74, Miller Puckette, 1997-2023)), un logiciel de programmation développé pour des applications dans les domaines de la musique et du multimédia. Il permet quatre types de manipulations sonores qui peuvent être combinées selon différentes configurations pour créer divers effets émotionnels. Les modifications de la tonalité émotionnelle de la voix sont ainsi produites par des algorithmes de traitement du signal audionumériques afin de simuler les caractéristiques acoustiques d'une émotion. Par exemple, pour la manipulation « joyeuse », la hauteur de la voix est modifiée avec un algorithme de décalage et d'inflexion de pitch pour la rendre plus positive, la plage dynamique de la voix est augmentée avec un algorithme de compression pour la rendre plus confiante, et son contenu spectral est modifié avec un filtre passe-haut pour la rendre plus brillante.

.

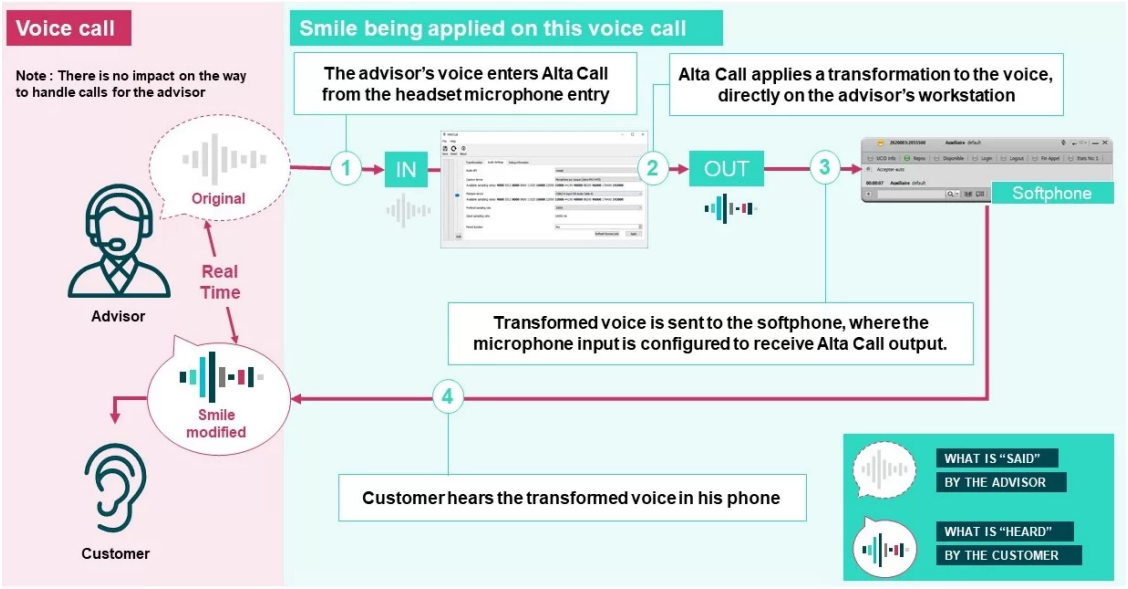


FIGURE 1.7 – Illustration de l'utilisation du filtre sourire au sein de call centers disponible sur le site de Webhelp https://thenest.webhelp.com/fr/how-can-voice-transformation-impact-your-csat-or-sales-rate.

En dehors du laboratoire, pouvons-nous imaginer une modalité d'usage de filtres de voix semblable à celle de leurs pendants visuels qui permettent notamment aux utilisateurs des plateformes de réseaux sociaux d'afficher à la demande et en temps réel un sourire algorithmique sur leurs selfies ? A notre connaissance, c'est au moins déjà le cas dans le secteur des call centers, où une transformation en temps réel (du type de celle décrite dans l'encadré ci-après) permettant d'attribuer une voix souriante aux employés est testée afin d'en évaluer le profit potentiel sur les ventes (voir Fig. 1.7). A cela s'ajoute toute une industrie de transformations vocales de loisir utilisées par les gamers pour prendre e.g. une voix de monstre dans les jeux vidéos : https://www.voicemod.net/discord-voice-changer/,https://voice.ai/apps/discord.

Exemple du filtre sourire de Arias et al. (2018a)

Plusieurs méthodes existent pour simuler l'effet de l'articulation sur les propriétés résonantes du conduit vocal sans modifier les autres aspects de la voix (comme sa f_0). Les plus simples, comme celle implémentée dans l'effet « change gender » du logiciel PRAAT Boersma (2001), exploitent les artefacts de la méthode de modification de f_0 par ré-échantillonnage qui augmente ou réduit l'écart entre toutes les résonances du conduit vocal. D'autres techniques permettent la modification spécifique de certaines bandes de fréquence (ou formants), par ex. les méthodes de resynthèse de formant ou de déformation d'enveloppe spectrale (frequency warping). Cette dernière méthode a par exemple été utilisée pour simuler l'effet du sourire sur la voix Arias et al. (2018b). Ainsi, pour construire cet effet, Arias et al. (2018a) ont enregistré un corpus de phonèmes français prononcés par des acteurs avec et sans sourire, et ont utilisé des techniques de transformation audio pour modéliser les changements observés dans l'enveloppe spectrale de ces sons. L'algorithme a augmenté la fréquence des deux premiers formants de l'enveloppe spectrale, et a augmenté l'amplitude du troisième formant. La transformation a été mise en œuvre à l'aide d'une architecture de vocodeur de phase (Liuni and Röbel, 2013), leur permettant ainsi de manipuler uniquement ces indices spécifiques au sourire, en laissant inchangées les autres caractéristiques de la voix, telles que la hauteur, le contenu, la vitesse et le sexe.

1.2.3.1 Effets du filtre vocal émotionnel sur les processus cognitifs : cas d'usage du paradigme de vocal feedback

A l'instar des filtres visuels, certains travaux en sciences cognitives nous invitent d'ores et déjà à penser les effets qui pourraient être associés à l'usage des filtres vocaux. C'est le cas des études utilisant le filtre vocal dans un paradigme expérimental particulier appelé rétroaction vocale émotionnelle (« *vocal feedback : VF* » en anglais) qui consiste à court-circuiter la perception habituelle de sa propre voix pour faire entendre à la place une voix à la tonalité émotionnelle fixée algorithmiquement.

Ainsi, l'usage de l'outil DAVID mentionné plus haut (Rachman et al., 2017) en condition de vocal feedback met en évidence un effet de « contagion » émotionnelle

avec l'émotion déterminée par le filtre (Aucouturier et al., 2016) (voir Fig. 1.8. Pour le dire autrement, l'évaluation subjective que les participants font de leur état émotionnel après modification par le dispositif va dans le sens de l'émotion créée par le filtre, alors même que les participants n'ont typiquement pas conscience de la modulation opérée (Goupil et al., 2021). Cette incidence du filtre vocal en condition de vocal feedback, qui a fait l'objet de multiples réplications notamment en langues anglaise, japonaise et suédoise (Goupil et al., 2021 ; Rachman et al., 2017) constitue un premier résultat en faveur d'une influence directe des filtres vocaux émotionnels sur le vécu subjectif réel des utilisateurs.

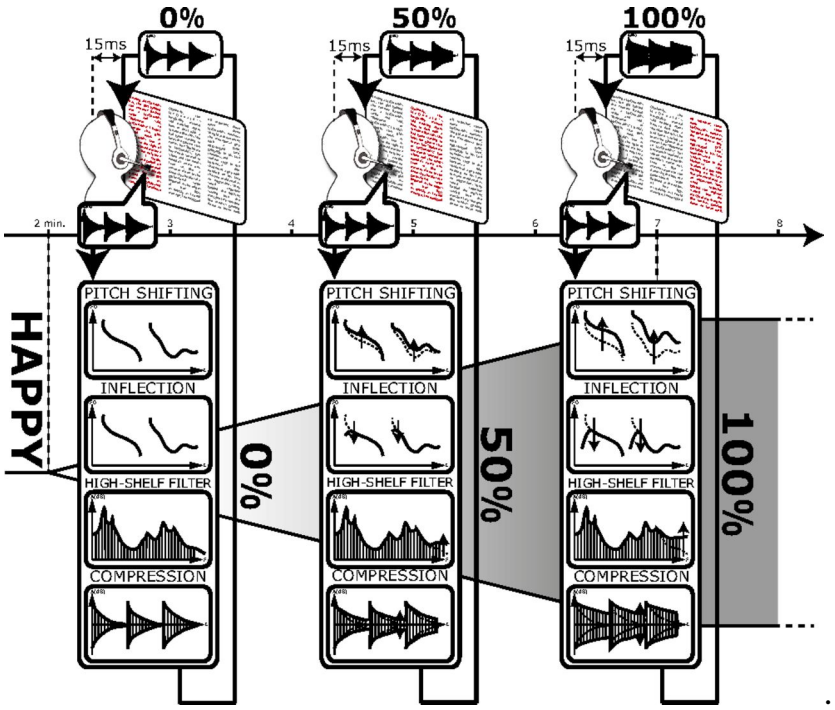


FIGURE 1.8 – Illustration du dispositif expérimental de rétroaction vocale émotionnelle tiré de l'article de Aucouturier et al. (2016)

Parmi les domaines qui se sont intéressés à la contagion émotionnelle sous vocal feedback figure celui de la recherche sur la gestion de conflits interpersonnels. En effet, si le vocable utilisé ainsi que le sujet du désaccord jouent un rôle important dans le processus de conflits, des recherches récentes ont révélé le rôle important des émotions comme déterminants de l'issue des conflits. Les travaux de Costa et al. (2018) étudient ainsi l'effet de filtres vocaux sur la régulation des émotions auprès

de couples engagés dans des interactions conflictuelles sur Skype (mode audio seul) en faisant écouter à un des deux partenaires via un casque d'écoute, sa voix dont le ton a été rendu plus calme par DAVID. Les résultats de leur étude ont montré qu'alors que l'anxiété des participants du groupe contrôle a augmenté après le conflit, les participants qui ont reçu un feedback vocal avec un ton plus calme se sont sentis moins anxieux.

Inscrivant ce résultat dans le cadre des théories de perception de soi, les auteurs proposent d'utiliser ce paradigme comme une méthode inédite de gestion des conflits en agissant sur la perception de sa propre voix.

1.2.3.2 Vers une technologie du soin : cas d'usage en psychiatrie ?

Ce potentiel de régulation du filtre vocal invite à imaginer divers contextes d'usage, parmi lesquels le domaine du soin émerge tout naturellement.

Sans même considérer les données de psychologie cognitive précédentes certains auteurs comme Godulla et al. (2021) posent qu'*avec l'amélioration constante de la technologie des deepfakes et sa prolifération actuelle, on peut s'attendre à ce que leur usage s'élargisse et en plus des contextes sociaux s'étende au domaine du soin.* De fait, l'extension de la cybersphère que nous vivons post-pandémie crée un environnement propice au sein duquel la digitalisation du soin opère déjà, l'émergence du cadre légal de la e-santé (Seguin and Tassy, 2022) témoignant de cette mutation opérée par le numérique dans ce domaine. Ainsi, selon le site de données statistiques publiques en santé et social, la pratique de la téléconsultation s'est généralisée en France : alors que moins de 5 % des médecins généralistes libéraux proposaient des consultations à distance à leurs patients avant l'épidémie de Covid-19, plus de trois sur quatre (77 %) indiquent début 2022 en avoir déjà effectuées. Il convient de souligner que les consultations ne nécessitant pas d'osculation et reposant majoritairement sur un entretien verbal entre soignant et patient se prêtent particulièrement à cette modalité.

Parmi les spécialités répondant à ces critères figure la psychiatrie. En effet, le diagnostic et l'évaluation de la sévérité des symptômes dans la maladie mentale reposent presque intégralement sur les entretiens cliniques avec le patient (Marmar et al., 2019). Dans ce cadre, la voix, qui est le support de ces échanges, porterait également les stigmates de la dérégulation émotionnelle qui caractérise nombre de désordres psychiatriques (Gross and Jazaieri, 2014). Comme nous le savons la

parole, outre qu'elle permet la transmission d'un langage signifiant, renseigne sur l'état émotionnel du locuteur (Banse and Scherer, 1996). A cet égard, plusieurs études ont relevé dans certains troubles psychiques une modification des facteurs extra linguistiques de la parole, c'est le cas par exemple dans la dépression majeure (Cannizzaro et al., 2004 ; Mundt et al., 2007 ; Yang et al., 2012)

Une des pathologies psychiatriques pour lesquelles le discours tient une place considérable est le Trouble de Stress Post Traumatique (TSPT). De fait, la thérapie gold standard indiquée pour un patient souffrant de TSPT - à savoir la thérapie d'exposition (ou réactivation) en imagination (Foa et al., 1991) qui sera l'objet de la section 3.1.4 - l'engage à raconter à voix haute et de manière très détaillée l'évènement traumatique à l'origine de la pathologie dont il souffre, ceci au cours de plusieurs séances consécutives. L'efficacité de la thérapie reposant en grande partie sur une diminution de l'intensité émotionnelle ressentie par le patient au cours de ces séances (Foa and Kozak, 1986) nous invite à considérer l'intérêt d'un filtre vocal manipulant la tonalité émotionnelle de sa voix en vue d'un effet facilitateur dans le processus de réduction de la charge émotionnelle associée au récit traumatique.

En de telles circonstances, la thérapie d'exposition en imagination pour le TSPT pourrait constituer à l'avenir un terrain propice pour l'application de technologies de manipulation de l'expressivité émotionnelle en psychiatrie. Dès lors, quelles seraient les caractéristiques du filtre vocal qui pourrait être utilisé pour cette thérapie et quels effets celui-ci aurait sur notre perception ? Notre thèse choisit donc d'examiner l'objet filtre vocal et son incidence sur les processus cognitifs en jeu en situation d'interaction en prenant pour cadre d'étude principal la thérapie d'exposition en imagination chez les patients souffrant de TSPT.

1.3 Cadre théorique de la perception des émotions

1.3.1 Externalisation de nos émotions : Clark & Chalmers

A la lumière des possibilités d'usages envisagées précédemment, il nous semble que ce que les filtres pourraient entraîner sur notre fonctionnement affectif serait de l'ordre d'une "externalisation", au même titre que celle conceptualisée au sujet de nos capacités cognitives par Clark and Chalmers (1998). L'externalisation de nos

capacités cognitives, en tant qu'objet d'étude intéressant à la fois les philosophes et les chercheurs en neurosciences cognitives, a donné lieu à différentes positions concernant les frontières plus ou moins étendues du siège de nos processus cognitifs. A ce propos, David Chalmers (philosophe connu pour avoir posé le "difficile problème de la conscience) et Andy Clark (philosophe et mathématicien) dans un article datant de 1998 tentaient de répondre à la question *« Where does the mind stop and the rest of the world begin ? »* et présentaient alors leur hypothèse d'une cognition étendue au corps et aux artefacts techniques. L'extension à la technique est celle qui nous intéresse tout particulièrement ici. A noter à la fin de leur article, la mise en avant des implications sociétales et morales qu'une telle externalisation pourrait porter avec elle. « *Les conséquences sont évidentes pour les conceptions philosophiques de l'esprit et pour la méthodologie de la recherche en sciences cognitives, mais il y aura aussi des implications dans les domaines moral et social. Il se peut, par exemple que dans certains cas, interférer avec l'environnement d'une personne aura la même signification morale que d'interférer avec sa personne* ». Dès lors, la question de l'acceptabilité morale de tels dispositifs se pose légitimement, question que nous aborderons au Chapitre 2 de cette thèse consacré à une étude d'éthique expérimentale au sujet de l'acceptabilité morale des filtres de transformations de l'émotion dans la voix.

Les filtres que nous envisageons dans le présent travail permettent un contrôle externe et artificiel de ce que nous donnons à entendre à nos interlocuteurs. Ces informations manipulées par informatique vont alors agir au niveau de la perception qu'ils ont de nous, perception qui est l'objet de notre questionnement. L'examen de ce dernier appelle de fait à étudier l'objet filtre à l'intérieur du cadre théorique précis de la perception dans le contexte des interactions sociales.

1.3.2 Perception et traitement prédictif : Friston & Frith

La perception met grandement en jeu des processus de déduction. Nous sommes à cet égard incapables d'accéder directement au contenu psychologique (ou à la psyche) de notre interlocuteur mais nous l'extrapolons à partir d'un flux continu d'informations que nous extrayons dans notre environnement et qui passent par le filtre de nos sens. Selon Friston and Frith (2015b) ce processus d'inférence relèverait de l'herméneutique. Classiquement, l'herméneutique renvoie à l'interprétation ou la traduction d'un texte

original en langue ancienne. En outre, elle renvoie également à l'interprétation dans le contexte de la communication verbale et constitue un cadre idéal de réflexion quant à la question de la lecture des émotions de notre interlocuteur (Friston and Frith, 2015b).

Jusqu'à récemment, la compréhension dominante du processus perceptif au sein des sciences cognitives se basait sur une conception du fonctionnement du cerveau pris tel un outil de traitement passif des informations. Ce modèle suggérait que le cerveau était conçu de manière à ce que les informations sensorielles provenant du monde (entrées ascendantes ou bottom-up) soient le principal moteur de la perception et de l'action. Ces informations sensorielles seraient ainsi relayées vers les régions corticales en charge des fonctions cognitives et perceptives dites de plus haut niveau et dictant en fin de compte l'action.

Les données récentes des neurosciences semblent nuancer ce propos. La perception ne reposerait pas principalement sur un phénomène passif mais au contraire sur un « *stratagème très simple mais remarquablement puissant : tenter de prédire les stimulations sensorielles en même temps qu'on les reçoit, en utilisant ce que nous savons à propos du monde* » (Clark, 2023).

Cette théorie dite de codage prédictif (Friston and Frith, 2015a) repose donc sur le postulat que la perception d'un stimulus nécessite que nous en prédisions la cause, et que cette opération de prédiction repose sur un modèle ou une heuristique du monde. Par exemple, pour percevoir la chute d'une pierre, nous faisons implicitement appel à un modèle de la manière dont les objets se déplacent sur Terre sous l'effet des forces gravitationnelles. Ce fonctionnement peut ainsi être étendu à la perception des états internes de nos interlocuteurs qui nous intéresse ici.

Deux des piliers de cette théorie du traitement prédictif (TP) sont les concepts de *prédiction* et de *modèle*, qu'il convient ici d'expliciter (Clark, 2015). Le terme prédiction dans son acception usuelle est une activité dans laquelle une personne s'engage afin d'anticiper la forme que pourraient prendre de futurs évènements. De telles prédictions consistent en des déductions *informées et conscientes*, souvent élaborées très en avance, et générées par des agents tournés vers l'avenir au service de leurs projets. Les prédictions au cœur de la théorie du TP sont d'une autre nature. Il s'agit d'un processus mis en œuvre automatiquement, de manière profonde ou hautement probabiliste et non consciente. De surcroît, il s'inscrit dans le cadre des routines

complexes de traitement neuronal qui sous-tendent et unifient la perception et l'action. Selon cette théorie, le cerveau est décrit comme maintenu dans une posture proactive continue. Il utiliserait un modèle génératif du monde - conçu et nourri à partir d'expériences passées - afin d'anticiper et répondre à des stimulations sensorielles nouvelles, tout en corrigeant (souvent mais pas toujours) les représentations internes du modèle dans le cas où celles-ci s'avéreraient inexactes (Clark, 2023).

Le processus d'inférence que nous venons d'expliciter est décrit comme reposant sur des probabilités bayésiennes (voir l'encadré sur l'inférence bayésienne). Ainsi, reformulé en termes informatiques : l'activité des neurones encode des croyances (*beliefs*, concrêtement des distributions de probabilité) sur les états du monde qui ont pu déclencher l'entrée sensorielle (par exemple : mes sensations visuelles sont provoquées par un visage). L'encodage le plus simple consiste à représenter la croyance par la valeur attendue ou la prévision d'une cause cachée. Ces causes sont dites cachées car elles doivent être déduites à partir des entrées sensorielles. Ce que fait alors essentiellement notre cerveau lors du processus de perception, c'est détecter des écarts par rapport à ses attentes, attentes générées par ses *modèles* statistiques du monde. Des estimations ratées génèrent des erreurs de prédiction (*prediction errors*, PE) qui sont utilisées pour recruter de nouvelles et meilleures estimations, ou pour informer des processus plus lents d'apprentissage et de plasticité. Ainsi, dans le cas où les PE se répètent, notre cerveau va chercher à modifier ses modèles du monde, donc à apprendre, grâce à la plasticité de ses connexions nerveuses (Bottemanne et al., 2022).

La théorie du TP permet ainsi de rendre compte de l'importance de l'environnement quant à son rôle de façonnement dans le phénomène de perception, et notamment celle des états internes de nos interlocuteurs lors de nos interactions sociales. D'une part, la notion de croyance rend compte de l'existence d'une « heuristique » qui se crée sur la base de l'expérience, souvent de manière non explicitement recherchée par l'observateur. Cette heuristique guide implicitement notre perception des états internes de nos interlocuteurs et peut concerner de manière attendue l'attribution d'émotions à celui-ci, ou encore d'un état interne plus complexe comme le désordre psychiatrique (*comment le clinicien apprend à percevoir les indices d'un possible TSPT chez son patient* - une question que nous aborderons au Chapitre 4). D'autre part, ces arguments nous invitent à penser que ce modèle génératif du monde pourrait

également prendre en charge des inférences au sujet des états physiologiques internes de nos interlocuteurs, les travaux de Galvez-Pol et al. (2022) mettant par exemple en évidence la capacité à détecter le rythme cardiaque de nos interlocuteurs à partir de vidéo de leurs visages - une question à laquelle nous reviendrons dans le Chapitre 5.

L'Inférence Bayésienne

Le cerveau réalise des inférences qui respectent le cadre mathématique de l'inférence bayésienne. L'inférence bayésienne est une théorie mathématique simple qui caractérise le raisonnement plausible en présence d'incertitudes. Elle rend compte des processus de perception qui reposent sur des entrées sensorielles ambiguës. Dans ce contexte d'ambiguité, notre cerveau en reconstruit l'interprétation la plus probable selon les modalités explicitées dans le théorème de Bayes.
La plausibilité de l'hypothèse $\mathcal{H}$, étant donnée une observation D, se définit selon le théorème de Bayes de cette manière :

$$p(\mathcal{H}|D) = p(D|\mathcal{H}).p(\mathcal{H})/p(D)$$

Dans cette équation, $p(\mathcal{H})$ est appelée probabilité *a priori* de $\mathcal{H}$ (prior en anglais). Notons à ce propos qu'il ne s'agit pas d'une connaissance rigoureusement indépendante de l'expérience, au contraire, l'a priori résulte souvent de données que nous aurions apprises d'expériences antérieures. Il s'agit en fait de la plausibilité de l'hypothèse avant de disposer des données D. En effet, sans avoir observé la moindre donnée, certaines hypothèses sont plus plausibles que d'autres dans notre environnement. $p(\mathcal{H}|D)$ est la probabilité a posteriori de $\mathcal{H}$. Ce terme n'est pas nécessairement à prendre dans son sens temporel, mais plutôt au sein d'une déduction logique : après avoir pris connaissance des données D, nous révisons la plausibilité de $\mathcal{H}$. Cette formule explicite comment réviser nos croyances. Enfin, $p(D|\mathcal{H})$, considérée comme une fonction de $\mathcal{H}$, est la vraisemblance de $\mathcal{H}$. Cette dernière ordonne les hypothèses $\mathcal{H}$ en fonction des observations et nous donne ainsi une idée des mérites relatifs des différentes hypothèses. Le facteur de Bayes qui sépare deux hypothèses ou deux « modèles » $\mathcal{M}_1$ et $\mathcal{M}_2$, est une mesure de leur mérite relatif, le rapport de leurs vraisemblances : $\mathcal{K} = p(D|\mathcal{M}_1)/p(D|\mathcal{M}_2)$. Ce facteur $\mathcal{K}$ renvoie à l' "évidence" en faveur d'une hypothèse par rapport à toutes les autres. Plus la valeur de $\mathcal{K}$ est élevée, plus on peut tirer des conséquences décisives sur le monde.
Cette équation pose donc en des termes simples et mathématiques l'influence des régularités statistiques de mon environnement sur la formation et la mise à jour de mes croyances sur le monde.

(Cet encadré est grandement inspiré du cours de S. Dehaene au Collège de France)

1.3.3 Emotions et Traitement Prédictif : Feldman Barrett & Searle

Ce paradigme de TP a récemment été repris et appliqué spécifiquement à la perception des émotions par la psychologue et neuroscientifique L. Feldman Barrett au sein de sa théorie constructiviste des émotions.

En dépit d'un ancrage notable de la théorie du TP au sein du connectivisme, la plupart des études en informatique émotionnelle continuent de fonder leurs travaux sur la théorie supposément darwinienne de l'universalisme des émotions et de leur caractère inné. A l'opposé, la théorie constructiviste de Barrett stipule que les émotions sont des artefacts culturels, des tentatives de donner un sens à nos entrées sensorielles façonnées par la culture dans laquelle nous évoluons. En guise d'exemple de cette construction sociale des émotions, Barrett (2012) note qu'il existe certaines régions du monde où la colère n'existe pas, c'est le cas notamment chez les Inuits Utku des Territoires du Nord-Ouest du Canada. Dire que la colère n'existe pas chez eux ne signifie pas qu'ils ne font pas l'expérience d'états semblables à ceux que nous vivons lorsque nous, occidentaux, déclarons être en colère. Pour autant, le sens qu'ils vont donner à cet ensemble de manifestations ne sera pas celui de « colère ». Les émotions seraient donc des constructions sociales qui vont donner du sens à ce que Barrett appelle *core affect* (Ekkekakis, 2013), c'est-à-dire une variation d'état sur les dimensions de valence et d'activation que l'on pourra différentiellement qualifier de peur ou d'excitation positive selon le contexte (situationnel ou culturel) dans lequel ces affects seront ressentis. Nous comprenons alors aisément en quoi cette théorie des émotions construites s'inscrit dans celle du TP, à savoir en mettant particulièrement l'accent sur le lien entre perception et connaissance du monde, les deux étant hautement interdépendantes (Gendron and Barrett, 2018).

Cette théorie des émotions, présentant il nous semble le phénomène de perception des émotions comme hautement créatif, est élaborée à partir des travaux de John Searle, philosophe américain. Searle (2010) pose que que les humains créent des objets ontologiquement subjectifs *au sein d'une réalité sociale* en assignant une fonction à ces objets qui n'est pas purement basée sur leurs propriétés physiques naturelles. Ce processus poïétique est explicité dans son œuvre comme une règle générale : un objet ou une instance (X) compte comme ayant un certain statut (Y) dans un contexte particulier (C). Ce statut permet à X d'acquérir une (ou des) fonction(s) particulière(s)

non inhérente(s) à sa structure physique. De cette manière, lorsqu'une plante est qualifiée de fleur ou de mauvaise herbe, cela confère une valeur à la plante : parler d'une plante comme d'une fleur signifie qu'elle doit être admirée et chérie, tandis que la considérer comme une mauvaise herbe la désigne comme quelque chose à rejeter. Les fleurs et les mauvaises herbes prescrivent des actions que les simples plantes ne peuvent pas faire. De ce fait, les fleurs doivent être cultivées et les mauvaises herbes doivent être arrachées du sol. Les plantes (X) deviennent des fleurs ou des mauvaises herbes (Y) dans un esprit humain (C) qui existe en cohérence avec d'autres esprits humains (comme c'est le cas dans une société donnée). Pour cela, ces esprits doivent être dotés de catégories pour les fleurs et les mauvaises herbes et être d'accord sur les fonctions de ces catégories.

Dans la continuité de Searle, la théorie constructiviste des émotions pose que les différents états corporels comme l'accélération du rythme cardiaque, une inflexion de voix ou une expression faciale existent bien, eux aussi, dans la nature mais, de la même manière que les plantes, ces états n'ont pas intrinsèquement de signification émotionnelle. Le fait que ces changements physiques soient reconnus comme manifestations de colère, de tristesse, de peur (ou même d'autres catégories psychologiques), relève du même processus que celui décrivant comment une plante devient une fleur ou une mauvaise herbe ; par l'intermédiaire de la cognition humaine qui confère à ces attribut physiques une signification émotionnelle. A travers ce sens qui leur est donné, les stimulations physiques acquièrent des fonctions particulières, qu'elles ne possédaient pas initialement, par elles-mêmes. Enfin, l'être humain selon Searle créerait des catégories ontologiques subjectives pour remplir des fonctions qui contribuent à la constitution de la vie sociale et la cimentent. En cela, nous pouvons considérer les catégories ontologiques subjectives émotions, comme des *outils cognitifs collectifs* qui permettent aux individus de la même culture d'associer un sens donné aux stimuli physiques de leur environnement. En appliquant la théorie du TP au contexte des émotions, le fait d'établir des catégories émotionnelles constituerait un mode de fonctionnement naturel du cerveau humain en tant qu'il essaie de donner du sens aux stimuli sensoriels ambigus de son environnement en utilisant les modèles de connaissances dont il dispose à son sujet. Parmi les situations ambiguës de notre quotidien, déchiffrer l'émotion dans la voix d'un interlocuteur apparaît exemplaire et obéirait alors également aux principes de

l'inférence bayésienne (voir Encadré 1.3.3). Le caractère émotionnel d'une voix ne relèverait donc pas uniquement de l'inné mais serait acquis par les régularités observées dans l'environnement et sous l'influence du modèle conçu à partir de ces régularités. Ce postulat théorique permet d'entrevoir l'influence de la culture et donc de ses objets (ici le filtre de voix) dans la perception des émotions (Bottemanne, 2021).

Encadré : Inférence bayesienne et perception des émotions

Afin d'exposer l'application des principes de l'inférence bayésienne à la perception des émotions, prenons l'exemple d'une situation d'interaction sociale qui se déroulerait entre une personne - que nous appellerons Paula pour rendre l'exposé plus fluide - et moi-même et décrivons la au moyen de l'équation de Bayes.

$$p(\mathcal{H}|D) = p(D|\mathcal{H}).p(\mathcal{H})/p(\mathcal{D})$$

Imaginons que je rencontre Paula dans les rues de Paris. Paula sourit (D). Parmi les multiples inférences (hypothèses) que je peux faire sur son état émotionnel à partir de ce sourire, prenons la suivante $\mathcal{H}$ = Paula est joyeuse. $p(\mathcal{H}/D)$ correspond ici à la plausibilité que Paula soit heureuse de me voir si au moment où je la croise dans la rue, je la vois sourire.
Comme nous l'avons vu, la probabilité de l'évènement prise séparément est importante dans le modèle pour juger de la probabilité de mon hypothèse a posteriori. Dans l'exemple que nous proposons, au sein de notre société occidentale, la probabilité de voir quelqu'un sourire est assez élevée $p(\mathcal{D})$ (en comparaison, par exemple, avec une société comme la Russie - Arapova (2016)). A contrario, voir quelqu'un pleurer dans la rue est assez rare .
De plus, la plausibilité de mon hypothèse dépend, selon la formule de Bayes, de la vraisemblance de la situation pour laquelle j'observe bien D quand $\mathcal{H}$ est vraie, ici voir Paula sourire alors qu'elle est heureuse de me voir. Cette vraisemblance $p(D/\mathcal{H})$ renvoie aux régularités observées et observables au sein d'un environnement précis. Pour notre exemple, jusqu'à présent, au cours d'une interaction mettant en jeu deux individus ayant grandi en France, l'observation de $\mathcal{D}$ lorsque $\mathcal{H}$ est vraie est assez fréquente et ne surprend personne. Nous pouvons donc supposer que le facteur $\mathcal{K}$ renvoyant à l'évidence en faveur de notre hypothèse qui est ici que Paula soit heureuse quand elle sourit par rapport à toutes les autres hypothèses que je pourrais faire au sujet de son état émotionnel, soit plutôt élevé. En conclusion, sur la base de notre exemple, il nous semble que l'on perçoit aisément la pertinence de ce cadre théorique pour mettre en évidence les processus d'inférence émotionnelle et tout particulièrement l'influence cruciale de l'environnement au sein duquel ces inférences sont faites.

1.3.4 Filtres de voix et potentiel de métamorphose des émotions

Ce caractère construit des émotions nous amène à nous questionner sur l'effet potentiel des filtres de voix car il nous semble que le processus à l'oeuvre relève de la « plasticité » au sens hégélien du terme comme le décrit la philosophe Catherine Malabou. En effet, la plasticité désigne à la fois la capacité de « *recevoir la forme (la terre glaise est plastique) et à donner la forme (comme dans les arts plastiques ou la chirurgie plastique)* » (Malabou, 2009). Il nous semble important d'appeler ici ce concept tant les idées de façonnement, modelage, transformation et d'adaptation qui lui sont attachées nous paraissent au coeur des processus qui intéressent notre thèse. Ainsi, de manière synthétique, nous pourrions dire que nos émotions sont plastiques et cette plasticité pourrait alors les rendre sensibles aux filtres vocaux. En effet, comme nous l'avons vu, les croyances composant nos modèles génératifs sont acquises en partie à partir de l'apprentissage d'associations entre stimuli récurrents au sein de notre environnement. Les nouvelles formes d'expression des émotions permises par les techniques de synthèse et de modification algorithmiques de la voix modifient les signaux perceptifs primaires présents dans notre environnement. Dès lors, elles pourraient agir sur nos modèles génératifs internes du monde.

Statistiquement, la culture occidentale nous enseigne, jour après jour, que le sourire est associé à des situations émotionnelles agréables (Rychlowska et al., 2015). Imaginons maintenant qu'il soit possible pour quiconque le souhaiterait de simuler une voix qui possède les caractéristiques jusqu'à présent associées à une émotion de « joie » et ceci de manière particulièrement convaincante dans n'importe quel contexte. Qu'adviendrait-il alors de ces régularités ? Serait-il toujours possible d'en trouver ? Si oui, qu'elles seraient-elles ? Nous réfléchissons à cette situation dans l'encadré suivant : *Application d'un filtre à l'inférence bayésienne*.

A la lumière de la théorie de l'inférence bayésienne, la question de la plasticité des émotions dans un monde avec filtre(s) apparaît alors se poser légitimement. Parmi les possibles, le caractère construit des émotions nous permet d'envisager que le statut d'émotion acquis par certains indices extéroceptifs comme la modulation de la voix, puisse sinon disparaître, au moins se modifier sur le long terme et par la même les fonctions que ce statut leur conférait (par exemple, relier le corps au monde, réguler l'action, la communication, l'influence sociale). Cela posé, notre travail s'intéresse

donc à l'influence des filtres de voix sur les processus cognitifs qui sous tendent nos inférences perceptives en situation d'interaction émotionnelle.

Encadré : application d'un filtre à l'inférence bayésienne

Reprenons maintenant le modèle de Bayes et son application en situation d'inférence émotionnelle que nous avons développé plus haut en l'illustrant par ma rencontre avec Paula dans les rues de Paris. Imaginons cette fois cette interaction au sein de notre écosystème numérique actuel : Paula et moi en réunion sur Zoom. Je me connecte sur la plateforme de visioconférence et Paula m'accueille avec un sourire. Faisons alors un petit effort d'imagination supplémentaire en nous figurant cette interaction dans un monde où nous disposons au sein de cet outil de communication de filtres émotionnels qui permettent de transformer en temps réel nos expressions émotionnelles.
Cette expérience de pensée nous permet d'envisager divers endroits d'incidence probables du filtre au sein de notre équation de Bayes. Ces hypothèses concernant l'incidence de l'usage du filtre sont à considérer sur un temps long, la récurrence des situations étant à la base des phénomènes d'apprentissage en jeu dans l'élaboration des croyances.

$$p(\mathcal{H}|D) = p(D|\mathcal{H}).p(\mathcal{H})/p(D)$$

Tout d'abord, le filtre permettrait un gain de contrôle inédit sur les expressions émotionnelles. Il pourrait alors agir sur la probabilité $p(D)$ de voir certaines expressions dans notre environnement. Dans le cas où certaines seraient jugées indésirables et qu'il devenait possible de les gommer complètement, cela modifierait nos priors à cet endroit. De surcroît, dans cette situation, l'utilisation d'un filtre permet de décorréler les deux éléments sourire (D) et l'état émotionnel "joie" de Paula ($\mathcal{H}$). Ceci pourrait avoir pour effet de perturber le mérite relatif (facteur $\mathcal{K}$) des différentes hypothèses que je peux faire à partir de ce sourire. De fait, le filtre permettrait de voir Paula sourire alors qu'elle est heureuse mais également lorsqu'elle ne l'est pas, alors même pourquoi pas qu'elle est complètement déprimée (augmentant la probabilité $p(D/\neg\mathcal{H})$. La valeur du K en faveur de l'hypothèse Paula est heureuse quand elle sourit pourrait alors diminuer. En conséquence, cela perturberait les régularités statistiques auxquelles nous sommes habitués et qui, nous l'avons vu plus haut sont cruciales à l'élaboration de nos connaissances sur le monde et à la mise à jour de nos a priori sur celui-ci étant donnée que nos croyances sont mises à jour en fonction de la vraisemblance attribuée à nos perceptions. Nous nous situerions alors dans un environnement plus incertain concernant les signes habituellement associées à nos états émotionnels, correspondant à une difficulté de stabiliser un modèle interne du monde.

1.4 Plan du manuscrit

C'est à l'intérieur de ce cadre que s'inscrit notre travail et, plus précisément, au moment de l'émergence d'un outil technique spécifique : le filtre vocal, capable de mimer des manifestations physiques que nous associons à des émotions. A la lumière des connaissances dont nous disposons concernant la perception des émotions et notamment du cadre théorique du traitement prédictif, nous posons la question de ce que l'utilisation de ce filtre vocal pourrait avoir comme conséquences sur les processus cognitifs guidant nos inférences au sujet des états de notre interlocuteur en situation émotionnelle. Même si, au moment de la rédaction de cette thèse, l'usage de l'objet filtre vocal n'est pas encore popularisé à la hauteur de son pendant visuel le *face filter*, de nombreuses études soulignent l'importance d'aborder ces questions bien en amont du déploiement général de ces technologies (Laakasuo et al., 2021). C'est ce projet que ce travail de thèse propose de mener à bien.

Considérant d'une part l'extension du numérique au soin, et de l'autre l'importance de la voix associée à l'intensité de l'expérience émotionnelle caractérisant toutes deux la situation de thérapie d'exposition en imagination pour le TSPT, nous avons choisi d'en faire le cadre de notre étude. De fait, le contexte de la thérapie d'exposition en imagination, qui implique de demander au patient de raconter son évènement traumatique à voix haute, nous permettra d'examiner les effets possibles du filtre vocal sur nos inférences perceptives en situation d'interaction émotionnelle.

Plusieurs niveaux d'observabilité seront alors abordés dans cette thèse. Tout d'abord, compte tenu des questions éthiques que d'utilisation d'un filtre de transformation de voix pose naturellement, nous avons souhaité commencer ce travail en interrogeant les formes d'usage de cet outil jugées acceptables par notre société, notamment dans le cadre thérapeutique. Aussi, nous présenterons au **Chapitre 2** une étude d'éthique expérimentale visant à interroger l'acceptabilité morale de l'usage de filtres permettant de modifier la tonalité émotionnelle de la voix et les conditions qui sous-tendent ce jugement.

Le **Chapitre 3** traite de l'importance de la voix dans le cadre spécifique de la thérapie d'exposition en imagination à destination des patients souffrant du TSPT. Cela posé, nous y présentons deux études longitudinales (étude TraumacoustiK et étude TraumacoustiK extension physiologique) menées dans cette population de

patients pour lesquelles nous analysons les corrélats acoustiques de l'évolution de leur symptomatologie au cours des séances successives de psychothérapie, investiguons leur lien avec un autre marqueur physiologique (l'activité cardiaque), et discutons la portée de ces données notamment sur le plan clinique.

Les résultats de ce chapitre au sujet des indices acoustiques extraits dans la voix d'une part et de leur lien avec le rythme cardiaque du locuteur de l'autre nous amènent ensuite à tester au sein des deux chapitres suivants l'incidence de la manipulation de la voix sur les processus d'inférences perceptives.

Ainsi, le **Chapitre** 4, dans la suite directe des résultats de TraumacoustiK, questionne les effets d'un filtre vocal - élaboré à partir des résultats acoustiques de l'étude clinique et appliqué à des extraits de voix de patients - sur les processus d'inférences perceptives des soignants au sujet des patients : les soignants sont-ils capables de percevoir dans la voix des indices de l'état de gravité du patient, et quel effet a le filtre de pitch sur ces inférences médicales ?

Enfin, au **Chapitre** 5, nous approfondissons le lien corrélationnel entre voix et activité cardiaque mis en évidence au sein de la deuxième étude clinique (TraumacoustiK extension physiologique) en l'examinant chez l'individu non malade en vue d'une plus grande maîtrise méthodologique. A cette fin, nous présentons dans une première partie une étude des effets sur la voix d'une manipulation causale du rythme cardiaque avec un protocole de tilt test. Puis nous examinons l'effet de la même transformation vocale qu'étudiée au Chapitre 3, cette fois sur les inférences qui peuvent être faites par un tiers au sujet du rythme cardiaque d'un locuteur à partir de sa voix.

En conclusion, nous faisons la synthèse de ce travail et tentons d'observer sa possible contribution à l'étude de notre question initiale : l'incidence de l'utilisation d'un filtre vocal sur nos inférences perceptives dans le cadre de l'interaction émotionnelle. En plus de nous pencher sur l'extension possible de cet outil de la perception d'autrui à celle de soi-même, nous discuterons des résultats de nos études au regard du potentiel de « transformation » humaine/sociétale que l'objet filtre semble porter et de son déploiement dans notre contexte sociétal actuel. Ce manuscrit se termine par une réflexion personnelle au sujet du positionnement de chercheur dans ce type de travaux, qui utilise les outils des sciences expérimentales pour interroger un objet technologique qui, de plus en plus, ne laisse pas la société indifférente.

2. Acceptabilité morale des filtres vocaux émotionnels

2.1 Introduction

Le contexte au sein duquel ce travail de thèse est réalisé est celui de l'émergence de technologies capables de contrôler et mimer artificiellement des signes expressifs extérieurs que nous associons généralement (en occident du moins) à des phénomènes émotionnels. Nous l'avons vu, ces technologies et leur espace de déploiement constitué par la cybersphère composent dorénavant avec différents domaines de nos existences et semblent susciter le débat. Afin d'entrevoir l'utilisation qui pourrait être faite de l'outil que nous étudions dans cette thèse, nous souhaitons dans un premier temps étudier son acceptabilité morale ainsi que les conditions qui la sous-tendent. De fait, ce chapitre traite donc de l'évaluation de l'acceptabilité morale de l'usage des filtres vocaux émotionnels dans divers scénarios suivant la méthodologie de l'éthique expérimentale.

Dans une étude menée en ligne, nous avons ainsi demandé à 303 jeunes français (majoritairement jeunes, urbains et étudiants) de lire 24 courtes vignettes décrivant chacune un scénario hypothétique d'utilisation de cette technologie de transformation de voix et d'évaluer l'acceptabilité morale de ces différents usages ainsi présentés. Afin de contextualiser cette évaluation, nous avons soumis aux participants une cover-story au sein de laquelle nous nous présentions comme une jeune start-up ayant développé un nouveau dispositif « MyVoicePlus » capable de transformer en temps réel les émotions dans la voix (voir Figure 2.1 pour illustration du dispositif) dont nous avions besoin de définir les futurs usages possibles en vue d'une commercialisation prochaine. Parmi les scénario présentés, on trouve par exemple celui du serveur de

café parisien qui transforme la voix de ses clients mal lunés pour moins subir leur agressivité, le cas d'un acteur qui gomme le trac dans sa voix, celui d'un patient dépressif rendant sa voix plus souriante lors de ses interactions sociales ou enfin d'une personne stressée qui se calme en écoutant sa propre voix rendue moins nerveuse. Pour créer les 24 scénario, nous avons fait varier quatre facteurs qui pourraient influencer l'acceptabilité de l'usage des filtres émotionnels :

1. le fait que l'utilisateur du filtre soit le participant ou un tiers inconnu,
2. le fait que le filtre soit utilisé dans un cadre thérapeutique ou à visée d'augmentation des capacités de l'utilisateur,
3. le fait que le filtre agisse sur des émotions plaisantes (augmenter le sourire) ou déplaisantes (atténuer la colère ou l'« anxiété »),
4. et le fait que le filtre agisse sur la manière dont la voix de l'utilisateur est entendue par autrui (càd la voix produite), sur les voix des interlocuteurs de l'utilisateur (voix perçue), ou alors sur la voix de l'utilisateur en situation de vocal-feedback, c'est à dire de modification de sa propre voix percevable par lui seul.

Pour chaque vignette, les participants ont d'abord évalué l'acceptabilité de la situation, puis ont été confrontés à deux dilemmes potentiels impliquant de mentir sur le véritable objectif de la transformation afin d'en améliorer l'efficacité. Enfin, pour tous ces jugements, nous avons examiné l'influence des différences individuelles au regard des attitudes morales des participants (Moral Foundations Questionnaire MFQ (Graham et al., 2011) et de leur familiarité avec la science-fiction (Science Fiction Hobbyism Scale SFH, (Laakasuo et al., 2018)), deux facteurs dont la pertinence a été montrée quant à la réception des nouvelles technologies (Körner et al., 2020 ; Koverola et al., 2020a ; Laakasuo et al., 2018 ; Medaglia et al., 2019).

Bien que notre étude soit exploratoire et que nous n'ayons pas formulé d'hypothèses formelles, un certain nombre de prédictions générales peuvent être faites à partir de la littérature sur la manière dont nos variables d'intérêt influenceraient les jugements moraux des participants. Premièrement, des expériences similaires portant sur des technologies émergentes telles que les véhicules autonomes (Bonnefon et al., 2016) ou la stimulation cérébrale (Medaglia et al., 2019) ont documenté des situations de dilemme social dans lesquelles les participants acceptent des choses pour eux-

mêmes (par exemple, une voiture qui favorise son conducteur plutôt que les piétons) qu'ils rejetteraient autrement pour les autres. Deuxièmement, pour diverses formes d'amélioration (mémoire, humeur, etc.), les participants sont généralement considérés comme étant plus à l'aise avec les technologies visant la récupération de capacités (c'est-à-dire qui sont utilisées à des fins thérapeutiques) plutôt que celles utilisées pour augmenter ces capacités au delà des limites de la norme (Cabrera et al., 2015 ; Koverola et al., 2020b). Enfin, à notre connaissance, il n'existe pas d'équivalent direct dans la littérature pour déterminer si, par exemple, la manipulation des émotions positives ou négatives, ou la manipulation de la perception ou de la production d'un utilisateur, a un impact sur le jugement d'acceptabilité des participants. La question de savoir si les participants se sentent plus à l'aise avec, par exemple, des filtres de sourire ou d'atténuation de l'anxiété, et des filtres qui manipulent la voix qu'ils produisent ou plutôt leur perception de la voix des autres, est une question ouverte qui nous paraît non triviale. A notre sens, en plus d'éclairer la réceptivité future du filtre vocal de transformation des émotions objet de notre thèse, elle nous renseignerait à un niveau plus général sur les futurs usages qui pourraient émerger en association avec cette nouvelle technologie.

2.2 Matériel et Méthodes

2.2.1 Participants

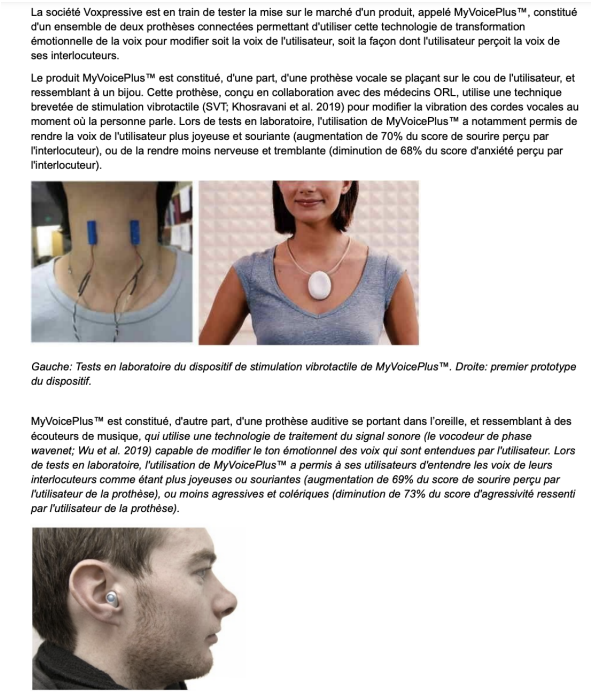
N=303 participants (M=25.7, femmes : 156) ont pris part à une étude en ligne administrée via la plateforme Qualtrics questionnaire (Qualtrics International Inc., Seattle, WA). Tous sont français, recrutés par le laboratoire comportemental INSEAD-Sorbonne Université parmi une population majoritairement consitituée d'étudiants à l'université. De fait, 213 participants (70.3%) avaient obtenu au moins le baccalauréat, et 116 (38%) au moins un diplôme de Master.
Les participants ont été répartis au hasard dans l'une des deux conditions « *soi* » ou « *autrui* ». Pour chaque condition, les participants se voyaient présenter 12 vignettes de scénarios évaluant trois facteurs internes testés pour leur impact possible sur l'acceptabilité morale (voir 2.2.3 pour le détail). Pour chaque vignette, les participants

ont répondu à trois questions sur leur perception de l'acceptabilité morale de la situation (voir 2.2.4), ce qui donne un total de 36 réponses pour chaque participant.

2.2.2 Procédure

Les participants se sont d'abord vu présenter une histoire décrivant un dispositif imaginaire capable de transformer la tonalité émotionnelle d'une voix en temps réel, à la fois dans la voix de l'utilisateur (modification de ce qu'il donne à entendre à ses interlocuteurs) et dans l'oreille de l'utilisateur (modification de la perception de la voix de ses interlocuteurs). Le dispositif, baptisé « *MyVoicePlus* », était présenté comme étant envisagé pour un éventuel déploiement commercial et/ou clinique par une jeune entreprise française. Cette cover-story comprenait des photographies fictives de l'appareil (composé à la fois d'une prothèse intra-auriculaire et d'une pièce au niveau du larynx, déguisée en bijou), ainsi que des références à la littérature technique sur la transformation de la voix (e.g., Khosravani et al. (2019)) décrivant prétendument les algorithmes implémentés dans le dispositif (voir Figure 2.1). Les participants ont été informés que la start-up commandait l'étude pour évaluer l'acceptabilité sociétale de leur technologie dans divers scénarios d'utilisation, et que leurs jugements collectifs conditionneraient le déploiement de la technologie.

Après lecture de la cover-story, on présentait aux participants une série de n=12 courtes vignettes -scénario. Le design de notre expérience comptait deux conditions inter-participants. Dans la condition *soi* (N=150), le participant lisait des vignettes le décrivant comme l'utilisateur de l'appareil (par exemple, « *imaginez que vous êtes un patient dépressif et que votre médecin vous conseille d'utiliser un dispositif de transformation de la voix... »*, alors qu'en condition *autrui* (N=153) les vignettes décrivaient des situations identiques mais dans lesquelles le dispositif était utilisé par d'autres personnes que le participant (par exemple, « *un patient dépressif... »*. Le participant était alors mis dans la position d'un interlocuteur de l'utilisateur. Pour chaque condition *soi* et *autrui*, les vignettes comprenaient un certain nombre de conditions intra-participant, que nous décrivons ci-dessous. Pour chaque vignette, les participants ont été invités à répondre à trois questions sur le degré d'acceptabilité morale de la situation selon eux (voir 2.2.4, plus bas).


La société Voxpressive est en train de tester la mise sur le marché d'un produit, appelé MyVoicePlus™, constitué d'un ensemble de deux prothèses connectées permettant d'utiliser cette technologie de transformation émotionnelle de la voix pour modifier soit la voix de l'utilisateur, soit la façon dont l'utilisateur perçoit la voix de ses interlocuteurs.

Le produit MyVoicePlus™ est constitué, d'une part, d'une prothèse vocale se plaçant sur le cou de l'utilisateur, et ressemblant à un bijou. Cette prothèse, conçu en collaboration avec des médecins ORL, utilise une technique brevetée de stimulation vibrotactile (SVT; Khosravani et al. 2019) pour modifier la vibration des cordes vocales au moment où la personne parle. Lors de tests en laboratoire, l'utilisation de MyVoicePlus™ a notamment permis de rendre la voix de l'utilisateur plus joyeuse et souriante (augmentation de 70% du score de sourire perçu par l'interlocuteur), ou de la rendre moins nerveuse et tremblante (diminution de 68% du score d'anxiété perçu par l'interlocuteur).

Gauche: Tests en laboratoire du dispositif de stimulation vibrotactile de MyVoicePlus™. Droite: premier prototype du dispositif.

MyVoicePlus™ est constitué, d'autre part, d'une prothèse auditive se portant dans l'oreille, et ressemblant à des écouteurs de musique, *qui utilise une technologie de traitement du signal sonore (le vocodeur de phase wavenet; Wu et al. 2019) capable de modifier le ton émotionnel des voix qui sont entendues par l'utilisateur. Lors de tests en laboratoire, l'utilisation de MyVoicePlus™ a permis à ses utilisateurs d'entendre les voix de leurs interlocuteurs comme étant plus joyeuses ou souriantes (augmentation de 69% du score de sourire perçu par l'utilisateur de la prothèse), ou moins agressives et colériques (diminution de 73% du score d'agressivité ressenti par l'utilisateur de la prothèse).*

FIGURE 2.1 – Capture d'écran de la présentation du dispositif *« MyVoicePlus » aux participants*

Enfin, après avoir répondu aux questions de toutes les vignettes, les participants étaient invités à remplir deux questionnaires standard mesurant les attitudes à l'égard de la moralité (Moral Foundations Questionnaire MFQ ; Graham et al. (2011)) et de la familiarité à la science-fiction (Science Fiction Hobbyism Scale ; Laakasuo et al. (2018)). L'étude durait environ 30 minutes.

2.2.3 Vignettes

Nous avons créé n=12 vignettes décrivant les applications potentielles du dispositif de transformation de la voix dans des situations concrètes de la vie quotidienne. Les vignettes variaient en fonction de trois facteurs situationnels décrits ci-après. Ils consitutent autant de variables intra-participants dont nous testerons l'influence sur l'acceptabilité du dispositif. Nous avons donc fait varier :

1. le contexte d'utilisation. A savoir si les transformations vocales sont utilisées pour restaurer (par exemple à des fins thérapeutiques) ou améliorer les capaci-

tés de l'utilisateur (condition *thérapie* : n=6 ; *augmentation* : n=6). Parmi les scénario d'usage thérapeutique on comptait l'utilisation d'un filtre par un patient dépressif afin de communiquer avec ses proches avec un ton de voix plus enthousiaste ou pour les scénario d'*augmentation* l'usage du même filtre par un président d'association afin d'améliorer son image auprès de ses membres. Dans les vignettes *thérapie*, le dispositif était décrit comme étant prescrit à l'utilisateur par un médecin alors que dans les vignettes *augmentation*, le dispositif était conseillé par un coach.

2. le type de transformation opérée par le filtre, atténuer la *colère* (n=2 ; par exemple rendre la voix des clients en colère moins pénible pour les opérateurs des centres d'appels), réduire l' *anxiété* (n=4 ; par exemple, aider un acteur en herbe à surmonter son trac) ou enfin augmenter le *sourire* (n=6).
3. la voix affectée par la transformation. A savoir si la transformation modifie la façon dont la voix de l'utilisateur est entendue par d'autres personnes (condition *production* : n=4), la façon dont l'utilisateur entend la voix d'autres personnes (condition *perception* : n=4), ou si elle est utilisée dans une situation où l'utilisateur entend sa propre voix manipulée (condition *feedback* : n=4). Parmi les exemples de conditions d'utilisation en vocal-feedback, nous avons présenté le fait de faire écouter à un patient souffrant d'un Trouble de Stress Post-Traumatique (TSPT) sa propre voix, rendue moins « anxieuse » pendant qu'il raconte son événement traumatisant. (Aucouturier et al., 2016).

La condition *soi – autrui* était attribuée au hasard entre les participants et toutes les autres conditions ont été randomisées aléatoirement au sein des participants. L'intégralité des vignettes est disponible dans l'article (Guerouaou et al., 2021).

2.2.4 Mesures :

Après avoir lu chaque vignette, les participants ont répondu à trois questions renseignant :

1. le niveau d'acceptabilité morale de la situation (*« A quel point jugez-vous cette utilisation du produit MyVoicePlus™ moralement acceptable ? »*)
2. le niveau d'acceptabilité d'utiliser le dispositif de manière cachée, c'est-à-dire sans informer ses interlocuteurs du fait que sa voix est modifiée ou que les voix

qu'il perçoit le sont, sachant que cela peut améliorer l'efficacité du dispositif jusqu'à 70 %. (« *A quel point jugez-vous acceptable le fait de cacher à votre entourage l'existence de la transformation de voix, en sachant que cela augmente considérablement l'efficacité du dispositif ?* »)

3. le niveau d'acceptabilité de cacher le véritable objectif de l'appareil à l'utilisateur, c'est-à-dire que les utilisateurs eux-mêmes ne savent pas qu'ils parlent ou entendent d'autres personnes avec une voix modifiée. (« *A quel point jugez-vous acceptable le fait que le médecin vous cache l'existence de la transformation de voix, en sachant que cela augmente considérablement son efficacité ?* »)

Les réponses aux trois questions ont été évaluées à l'aide d'une échelle de Likert en 9 points allant de 1 « *complètement inacceptable* » à 9 « *complètement acceptable* ».

2.2.5 Questionnaires

Outre les jugements moraux portés sur les vignettes, les participants ont rempli deux questionnaires mesurant leurs attitudes à l'égard de la moralité (Moral Foundations Questionnaire MFQ ; Graham et al. (2011)) et à l'égard de la technologie et de la science-fiction (Science Fiction Hobbyism Scale SFH , Laakasuo et al. (2018)).

La version française du MFQ utilisée ici (Métayer and Pahlavan, 2014) consiste en 32 questions courtes portant sur la pertinence de diverses considérations (par exemple, le fait de savoir si une personne a souffert émotionnellement) lorsqu'il s'agit de décider si une chose est bonne ou mauvaise, avec une note allant de 1 (*pas du tout pertinent*) à 7 (*extrêmement pertinent*) et le degré d'accord du participant avec diverses positions morales (par exemple : « *La compassion pour ceux qui souffrent est la vertu la plus cruciale* » ; noté de 1 (*très en désaccord*) à 7 (*très en accord*). Conformément à l'analyse classique du MFQ (Graham et al., 2011), nous avons regroupé et calculé la moyenne des réponses de chaque participant sur les cinq sous-échelles suivantes : soins - préjudice (6 items ; par exemple, « *si quelqu'un a souffert émotionnellement* »), équité - tricherie (5 items ; par exemple « *si certaines personnes ont été traitées différemment des autres* »), loyauté - trahison (6 items ; par exemple « *si quelqu'un a fait quelque chose pour trahir son groupe* »), autorité - subversion (5 items ; par exemple « *si une action a provoqué le chaos ou le désordre* »), et pureté - dégradation (6 items ; par exemple « *si quelqu'un a violé les normes de pureté et de décence* ».

L'échelle SFH (Laakasuo et al., 2018) est composée de 12 items et mesure l'exposition culturelle des individus aux technologies futuristes et aux thèmes de science-fiction (exemples d'items : « *Je pense souvent à ce que seront les machines dans le futur* », « *Je repère souvent des erreurs liées à la science ou à la technologie dans les films, les séries télévisées ou les livres de science-fiction* »). Toutes les questions sont notées de 1 (pas du tout d'accord) à 7 (tout à fait d'accord), les scores les plus élevés indiquant une plus grande familiarité avec la science-fiction. Une étude antérieure (Koverola et al., 2020a) a validé l'échelle sur N=172 participants et a montré qu'elle avait de bonnes propriétés psychométriques (tous les facteurs $> .57$; α de Cronbach=.92). Dans ce travail, nous avons utilisé notre propre traduction française non validée du SFH, pour laquelle nos données (N=303) ont révélé une bonne validité interne (α=0.89, [0.888, 0.898]).

2.2.6 Analyses statistiques

L'étude comportait deux variables dépendantes (VD), mesurant l'acceptabilité de l'utilisation transparente (VD1) et dissimulée (VD2) des transformations vocales. La VD2 a été construite en recodant les deux questions sur l'utilisation dissimulée (mensonge à l'utilisateur et mensonge aux autres) en une seule VD mesurée dans deux conditions (à qui le filtre est caché).

Les vignettes de l'étude couvraient un certain nombre de caractéristiques situationnelles, chacune décrite comme une combinaison de variables indépendantes (VI). L'expérience compte ainsi une VI inter-participants (soi–autrui), trois VI intra-participants pour la VD1 (restauration – augmentation, sourire – anxiété – colère, production – perception – feedback) et deux VI intra-participants supplémentaires pour la VD2 (mensonge à l'utilisateur – autrui, et mensonge au participant – autrui). L'analyse de l'effet des VI sur les deux VD a été faite à l'aide d'ANOVA, à mesures répétées ou mixtes, en calculant la moyenne des scores d'acceptabilité intra-participant sur les vignettes correspondant à chaque condition testée.

En outre, 6 mesures de caractéristiques individuelles ont été également relevées (MFQ : 5 concepts ; SFH : 1 concept). Pour chacune d'elles, nous avons testé l'association avec les VD de l'étude en calculant les moyennes des scores d'acceptabilité en intra-participants (un point de données par participant) et en procédant à une

régression multiple. Nous avons testé la condition d'hétéroscedasticité des résidus à l'aide du test de Breusch-Pagan. En cas d'homoscedasticité, nous avons utilisé des régressions multiples en utilisant la méthode des moindres carrés (ordinary least square –OLS) ; en cas d'hétéroscedasticité nous avons utilisé des régressions de type re-weighted least-square (IRLS) associées à la correction de Huber .

Toutes les analyses ont été conduites en Python (3.6.8), en utilisant les libraires `seaborn` (© M. Waskom, 2021-2023) et `Pingouin` (© R. Vallat, 2018-2023). Les données ainsi que le code des analyses sont disponibles ici https://github.com/creamlab/deep-ethics.

2.2.7 Réglementation éthique

Tous les participants ont été testés au Centre des sciences du comportement de la Sorbonne et de l'INSEAD. L'expérience a été approuvée par l'IRB de l'Institut Européen d'Administration des Affaires (INSEAD) (Étude 202058 ; « Étude de l'acceptabilité morale à l'égard de l'utilisation d'un dispositif de transformation de la voix » ; décision du 18 juin 2020). Tous les participants ont donné leur consentement éclairé pour l'étude et ont été indemnisés pour leur participation.

2.3 Résultats

2.3.1 Acceptabilité de l'usage non caché du filtre

Dans un premier temps nous avons évalué l'acceptabilité morale de l'utilisation d'un dispositif de transformation de voix jugée par les participants (N=303) dans le cas où la transformation est clairement explicitée à toutes les parties.

2.3.1.1 Les filtres vocaux sont généralement bien acceptés par la population

Parmi les différents situations, l'acceptabilité morale des transformations de voix utilisée de manière explicite était largement significativement supérieure au jugement neutre (M=6.49>5 ; t-test unilatéral par rapport au point médian, en moyennant les scores d'acceptabilité de toutes les vignettes à la moyenne : $t(302)=146$, $p<.001$).

Compte tenu de l'hétéroscedasticité de nos données, (Breush-Pagan : F(6,296)= 3.23, p=.004), nous avons testé l'effet des caractéristiques individuelles sur ce jugement au moyen de régressions multiples IRLS. L'acceptabilité était positivement associée de façon significative avec la familiarité des participants à la science-fiction (β=0.014, z=2.75, p=.006 ; Figure 2.2-C) et marginalement positivement associée au niveau de MFQ-pureté (MFQ-PU ; β=0.04, z=1.85, p=.064). Aucun autre facteur de MFQ ne montrait de régression significative (tous les $p>$ 0.1). Cette dernière association avec MFQ-PU diffère des résultats retrouvés dans d'autres études concernant des technologies d'augmentation pour lesquelles la pureté était plutôt corrélée de manière négative avec l'acceptabilité (voir mind upload Laakasuo et al. (2018) ; robots sexuels Koverola et al. (2020b)).

2.3.1.2 Une utilisation en contexte thérapeutique rend l'usage des filtres vocaux d'autant mieux accepté

Nous avons testé l'effet de l'objectif sous-jacent à l'usage des filtres vocaux, à savoir restaurer ou augmenter les capacités de l'utilisateur sur l'acceptabilité morale des filtres utilisés en condition transparente. Pour cela nous avons moyenné en intra-participants les scores d'acceptabilité concernant les scénario d'usage thérapeutique (n=6 vignettes) et d'augmentation (n=6 vignettes), et avons testé l'effet de cette condition avec une ANOVA univariée en mesures répétées. Le facteur réparation – augmentation présentait un effet principal significatif sur l'acceptabilité de la situation (F(1,302)=47, $p<$.001), les usages thérapeutiques étant encore mieux acceptés (M=6.7) que ceux visant l'augmentation des capacités (M=6.2 ; Figure 2.2-A).

Afin de savoir si cette effet du facteur réparation – augmentation était associé à des caractéristiques individuelles, nous avons calculé la différence moyenne intra-participant entre les scores d'acceptabilité calculés sur les deux types de vignettes et appliqué une régression multiple (OLS) (test d'hétéroscédasticité de Breusch-Pagan : F(6,296)=0.39, p=0.88). La meilleure acceptabilité de l'usage thérapeutique des filtres vocaux n'était associé à aucune différence individuelle concernant la familiarité à la science-fiction ni au MFQ (R^2 = 0.008, F(6,296)=0.38, p=.89).

2.3.1.3 Manipuler la perception est moins acceptable que manipuler la production

De la même manière, nous avons testé si le fait que le filtre joue sur la voix produite par l'utilisateur (modification visant sa propre voix et qui sera perçue par son interlocuteur ; condition *production* : n=4 vignettes), sur les voix perçues par l'utilisateur (modification de la voix de ses interlocuteurs ;condition *perception* : n=4 vignettes), ou en condition de vocal feedback (modification visant la voix de l'utilisateur perçue par lui-même ; condition *feedback* : n=4 vignettes). Pour cela, nous moyennons les scores d'acceptabilité intra-participants correspondant aux trois types de situations et testons l'effet de cette variable sur les jugements d'acceptabilité au moyen d'une ANOVA univariée à mesures répétées. Nous avons observé un effet significatif de la variable production – perception – feedback variable (F(2,604)=7.5, p=.001), les transformations affectant la voix produite par les utilisateurs étant mieux acceptées que celles visant les voix qu'il perçoit (qu'il s'agisse de celle de ses interlocuteur ou la sienne).
Nous avons testé l'influence des caractéristiques interindividuelles sur cet effet en calculant la différence moyenne d'acceptabilité entre les trois conditions et avons calculé une régression multiple OLS sur ces valeurs (test d'hétéroscédasticité de Breusch-Pagan : perception - production F(6,296)=0.27, p=0.96 ; feedback - production F(6,296)=0.66, p=0.6815). La différence d'acceptabilité entre ces conditions n'était associée ni aux scores de MFQ ni de SFH des participants (perception - production : R^2 = 0.006, F(6,296)=0.31, p=.93 ; feedback - production : R^2 = 0.011, F(6,296)=0.571, p=.75).

2.3.2 Acceptabilité des usages cachés

Pour chaque situation, nous avons ensuite testé l'acceptabilité du fait de cacher la vraie transformation opérée par le filtre afin d'augmenter l'efficacité de la transformation de voix, dans deux situations qui impliquait pour l'une que l'utilisateur du filtre mente à ses interlocuteurs et pour l'autre que le prescripteur de la transformation mente à l'utilisateur. Sachant que l'utilisation de ces filtres vocaux s'avère être généralement bien acceptée (voir plus haut), et parce que nous avons justifié l'usage caché du filtre par une meilleure efficacité, ces situations peuvent être considérées comme

de véritables dilemmes moraux pour lesquels une action déontologiquement répréhensible pourrait être justifiée par la valeur utilitaire de l'amélioration des performances qui en résulte.

2.3.2.1 Utiliser les filtres vocaux de manière cachée n'est pas un problème...

Bien que les situations jugées les plus acceptables présentaient également une meilleure acceptabilité pour un usage caché,(régression OLS sur les jugements moyens inter participants : R^2=0.61, F(1,22)= 34.61, p <.001 ; test d'hétéroscédasticité de Breusch-Pagan : F(1,22)=0.13, p=0.72), mentir au sujet de la transformation opérée par le filtre était généralement considéré comme non acceptable (M=4.69 $<$ 5, t(302)=3.19, p=.001).
Cependant, nos résultats montrent une interaction importante avec l'identité de la personne à laquelle la transformation est cachée (ANOVA univariée à mesures répétées : F(1,302)=631, p <.001 ; Figure 2.2-D) : de manière surprenante, les participants considèrent moralement acceptable le fait que l'utilisateur du filtre vocal cache la transformation opérée sur sa voix à ses interlocuteurs (M= 6.08 [5.87,6.3] ; t-test unilatéral par rapport au point médian : t(302)=9.99, p <.001).
Compte tenu d'une hétéroscedasticité marginale (Breusch-Pagan : F(6,296)=1.72, p=0.11), nous avons testé l'association entre l'acceptabilité de la situation au sein de laquelle l'utilisateur du filtre ment à ses interlocuteurs et les caractéristiques individuelles des participants au moyen de régressions multiples IRLS. L'acceptabilité d'une utilisation cachée à ses interlocuteurs ne présentait aucune association avec les différents sous scores de la MFQ (meilleur PU : β=0.033,z=1.33, p=.18), mais était associée de manière positive avec la familiarité à la science-fiction (β=0.01, z=1.95, p=.05).

2.3.2.2 ...à moins que l'on ne le cache à son utilisateur

Cependant, cacher la transformation à la personne utilisant le filtre vocal est jugée comme totalement inacceptable (M=3.3<<5, t test à une condition par rapport au point médian t(302)=-14.9, p <.001)(Figure 2.2-D).
De la même manière que précédemment, nous avons testé l'association de l'acceptabilité des situations dans lesquelles la transformation était cachée à l'utilisateur avec les

caractéristiques individuelles des participants en utilisant des régressions multiples IRLS (test d'hétéroscédasticité de Breusch-Pagan : $F(6,296)=1.28$, $p=0.26$). La faible acceptabilité du fait de cacher la transformation à l'utilisateur était négativement associée à des scores élevés à la sous dimension de justice du MFQ (β=-0.1395 ; z=-3.461, p=.001 ; Figure 2.2-E) mais atténuée (i.e. positivement associée) par de hauts scores de pureté (β=0.0930, z=3.65, p <.001 ; Figure 2.2-F) et de loyauté au MFQ (β=0.09, z=2.71, p=.007). L'acceptabilité d'une transformation dont l'effet serait caché à l'utilisateur semble également associée à la familiarité à la science-fiction (β=0.014, z=2.38, p=.017), une plus grande familiarité étant liée à des scores d'acceptabilité plus élevés.

2.3.3 L'acceptabilité des filtres vocaux n'est pas influencée par la recherche de profit personnel

Nous avons testé l'effet de première personne dans nos scénarios, à savoir si les situations décrites mettaient le participant dans la peau de l'utilisateur (utilisation du pronom vous) ou alors si les scénarios décrits relataient les usages d'un tiers (un garçon de café, un patient dépressif...). Pour cela, nous avons réalisé des ANOVA mixtes, en posant la variable soi–autrui comme facteur inter-participants, et les différentes conditions décrites par les vignettes (réparer – augmenter, transformation positive – négative, production – perception – feedback) comme facteurs intra-participants. En condition d'utilisation explicite des filtres (usage du dispositif connu de toutes les parties), aucune différence significative n'était observée entre les scénarios décrivant un usage fait par le participant (M=6.43), ou par un tiers inconnu (M=6.55 ; pas d'effet principal, $F(1,301)=0.37$, p=.54). Aucune autre interaction n'a été retrouvée pour le facteur soi–autrui avec les différentes conditions expérimentales. De fait ce facteur n'a aucune influence sur les différences entre usages à visée thérapeutique ou d'augmentation (pas d'interaction entre soi–autrui $\times$ réparer – augmenter : $F(1,301)=1.68$, p=.20 ; Figure 2.2-A), les différences entre les transformations visant le sourire, l'anxiété ou la colère dans la voix (pas d'interaction soi – autrui $\times$ transformation : $F(1,301)=1.43$, p=.23 ; Figure 2.2-B), et enfin sur les différences observées entre les situations de production, perception ou feedback (pas d'interaction soi – autrui $\times$ production – perception – fb : $F(2,602)=0.047$, p=.95).

De la même manière, les participants n'ont pas jugé moins acceptables les situations dans lesquelles la transformation de voix leur était cachée (qu'ils soient l'utilisateur ou non) par rapport à des situations pour lesquelles la transformation était cachée à un tiers inconnu (ANOVA-mr, avec pour facteur intra participant le point de vue du participant soi – autrui, $F(1,302) = 0.0026$, $p = .87$; Figure 2.2-D). En d'autres termes, la relativement grande acceptabilité de l'utilisation cachée des filtres vocaux ne semble pas dépendre du fait que le participant en soit l'utilisateur ou plutôt la personne à laquelle à la transformation est cachée. Aussi, la faible acceptabilité du fait de mentir à l'utilisateur du filtre ne dépend pas non plus du fait que le participant soit mis dans la peau de l'utilisateur ou non.

En résumé, contrairement à certaines situations décrites dans la littérature comme dans le cas de l'usage des véhicules autonomes (Bonnefon et al., 2016), il ne semble pas y avoir de dilemme social concernant l'utilisation de filtres vocaux, dilemme qui impliquerait que l'on juge acceptables pour soi-même une conduite que l'on ne tolérerait pas chez autrui.

2.3.4 La nature de l'émotion impacte l'acceptabilité morale de la transformation

Enfin, nous avons testé l'influence de la nature de l'émotion visée par le filtre vocal sur l'acceptabilité de la situation ainsi que l'interaction avec le contexte d'utilisation thérapie–augmentation). Pour ce faire, nous avons moyenné les scores en intra-participants de jugements d'acceptabilité dans un cadre d'usage explicite pour les vignettes illustrant des transformations en conditions thérapeutique et d'augmentation visant les émotions anxiété (n=4 vignettes ; thérapie :2), colère (n=2 vignettes ; thérapie :1) et sourire (n=6 ; thérapie : 3), et avons réalisé une ANOVA-rm à deux facteurs.

Les résultats indiquent un effet principal de l'émotion : les situations pour lesquelles le filtre visait à atténuer les deux émotions de valence négative anxiété (M=6.8) et colère (M=6.5) étaient mieux acceptées que celles décrivant une augmentation du sourire ($F(2,604) = 24.47$, $p < .001$), même si ces dernières restent quand même bien acceptées (Figure 2.2-B).

Cet effet de l'émotion interagit significativement avec la variable réparer – augmenter (F(2,604)=21.3, p <.001), les transformations visant à atténuer les émotions déplaisantes sont jugées encore plus acceptables en condition thérapeutique (Δ=+0.56) que les transformations visant le sourire (Δ=+0.35). Dès lors, l'effet maximal est retrouvé pour l'atténuation de l'anxiété dans la voix à visée thérapeutique (réparer : M=7.17 ; augmenter : M=6.34).
De la même manière, dans les situations pour lesquelles la transformation était cachée, il semblait plus acceptable de dissimuler le véritable effet du filtre lorsque celui-ci visait à atténuer les émotions déplaisantes que lorsqu'il ciblait le sourire (ANOVA-rm à un facteur ; effet principal de transformation : F(2,604)=8.3, p <.001).
Enfin, nous avons testé si ces différences entre transformations visant à atténuer des émotions de valence négative ou augmenter le sourire étaient associées à des différences individuelles chez les participants en calculant les différences moyennes en intra-participants entre les trois types de transformations sur lesquelles nous avons appliqué une régression multiple OLS (test d'hétéroscédasticité de Breusch-Pagan : anxiété - sourire F(6,296)=0.59, p=0.74 ; colère - sourire F(6,296)=0.73, p=0.62). La différence d'acceptabilité entre ces deux situations n'était associée ni au MFQ, ni à la SFH (anxiété - sourire : R^2 = 0.031, F(6,296)=1.58, p=.15 ; colère - sourire : R^2 = 0.023, F(6,296)=1.137, p=.34).

2.4 Discussion

Nous présentons ici une étude d'éthique expérimentale dans laquelle N=303 participants en ligne ont évalué l'acceptabilité de vignettes décrivant des applications potentielles de filtre de transformation des émotions dans la voix. Nos résultats indiquent que ces filtres étaient généralement bien acceptés, notamment dans un contexte thérapeutique (par opposition à celui d'augmentation), lorsqu'ils visaient à atténuer des émotions de valence négatives plutôt que d'intensifier des émotions positives, et lorsqu'ils manipulaient la production d'un locuteur plutôt que sa perception. De manière surprenante, les transformations vocales sont restées bien acceptées même lorsque l'utilisateur cachait à ses interlocuteurs leur utilisation et, contrairement à d'autres technologies émergentes telles que les véhicules autonomes, nous n'avons pas mis en évidence de dilemme social dans lequel on accepterait, par exemple, pour

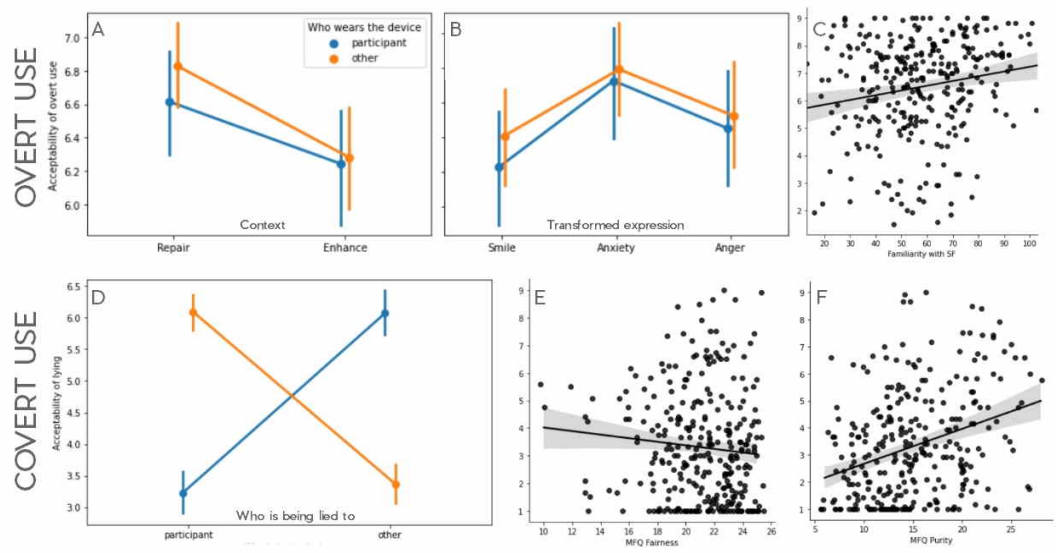


FIGURE 2.2 – Jugement d'acceptabilité pour les usages explicite et caché des filtres vocaux. Top : Usage explicite. (A) L'acceptabilité morale en condition explicite est plus haute que le point médian, et les transformations en situation thérapeutiques sont encore mieux acceptées que celles visant l'augmentation. (B) Les situations pour lesquelles les transformations visent à atténuer les émotions de valence négative d'anxiété ou de colère dans la voix étaient mieux acceptées que les filtres permettant d'augmenter le sourire dans la voix. (C) Dans toutes les situations, l'acceptabilité est positivement associée à la familiarité des participants à la Science-Fiction. Bas : Usage caché impliquant de mentir au sujet de l'objet réel de la transformation vocale afin d'en augmenter l'efficacité. (D) Les participants considèrent moralement acceptable le fait que l'utilisateur de la transformation de voix cache son objectif réel à ses interlocuteurs mais cacher la réelle transformation opérée par le filtre vocal à son utilisateur est complètement inacceptable. (E-D) L'acceptabilité du fait de cacher la transformation à l'utilisateur du filtre était négativement associée au souci des participants au regard des valeurs de justice, mais positivement liée à la sous échelle de pureté du MFQ.(A,B,D) Pour toutes les conditions, aucun effet du fait que l'utilisateur du filtre soit présenté dans le scénario comme étant le participant lui-même ou une tierce personne, barres d'erreur : 95% CI.

les autres ce qu'on ne veut pas pour soi-même. Nous posons que le fait de ne pas observer ici de jugements intéressés pourrait témoigner de l'absence de méfiance des participants à l'égard de cette technologie. En effet, de récentes études de Weiss et al. (2018) montrent que les participants qui ont tendance à éprouver de la méfiance (par rapport à ceux qui ont confiance) adoptent des normes morales plus indulgentes envers eux-mêmes que pour les autres. Au sein de nos données, la seule véritable objection morale à ces filtres vocaux semble liée aux situations dans lesquelles ils sont

appliqués à un locuteur à son insu, l'acceptabilité de telles situations étant modulée par les différences individuelles en matière de valeurs morales et de familiarité à la science-fiction. Ces différents résultats sont discutés en détail dans l'article (Guerouaou et al., 2021), néanmoins nous limiterons ici l'examen des points qui intéressent directement le cas de l'utilisation du filtre vocal qui fait l'objet de la thèse.

Le fait que les transformations vocales soient généralement bien acceptées montre avant tout que la population occidentale, jeune et éduquée étudiée ici est favorable à l'idée de personnaliser sa propre expression émotionnelle grâce à une technologie de transformation de voix.

Aussi, le fait que les filtres de voix soient encore mieux acceptés en situation d'usage thérapeutique par rapport à l'utilisation à des fins d'augmentation (Kass, 2003) est cohérent avec ce qui est rapporté dans d'autres études relatives à l'augmentation des capacités cognitives (Cabrera et al., 2015 ; Koverola et al., 2020b ; Medaglia et al., 2019) d'une part, et avec les impératifs mis en avant par la littérature bioéthique (Persson and Savulescu, 2008 ; Sahakian and Morein-Zamir, 2011) d'autre part. Dès lors, ce résultat confirmerait que la distinction entre thérapie et augmentation est moralement importante pour le public concernant l'usage potentiel d'une technologie de transformation de l'émotion dans la voix.

En plus d'avoir manipulé le contexte d'usage de ces filtres, nous nous sommes intéressés à la question de l'importance des processus cognitifs visés par les transformations de voix. A ce propos, nous avons constaté une acceptation encore plus grande pour les filtres qui transforment la production vocale que pour ceux qui manipulent sa perception. Le fait que les participants semblent avoir un préjugé défavorable à l'égard de ces dernières contredit en quelque sorte l'attente selon laquelle des changements proprement internes à l'individu seraient mieux acceptés que des transformations affectant son expression extérieure (Kahane, 2015). Cette préférence pourrait refléter une crainte de voir son expérience réelle déformée. Néanmoins cette hypothèse explicative se heurte à l'absence d'effet soi–autrui (les situations en conditions autrui plaçant le participant dans la position d'un interlocuteur de l'utilisateur du filtre, donc pour les vignettes production–autrui, sujet à une distorsion de ses perceptions). Une autre hypothèse pourrait tenir de la familiarité actuelle aux filtres visuels qui sont en premier lieu pensés comme des objets visant à modifier ce que l'on donne à voir à autrui, comparé à des dispositifs tels que des lunettes de réalité augmentée (AR)

dont l'usage reste beaucoup plus limité au moment de la rédaction de ce manuscrit. De fait, il serait d'examiner la persistance de cette différence à l'avenir si ces derniers dispositifs venaient à gagner en popularité (Michaud, 2020).

Parallèlement aux processus cognitifs visés par le filtre, nos résultats montrent un effet du type d'émotion visée par la transformation, l'atténuation des émotions de valence négative (anxiété et colère) étant encore mieux acceptée que l'amplification du sourire dans la voix. Cette asymétrie contraste avec une vision purement hédoniste pour laquelle agir de telle sorte que les individus se sentent « aussi bien que possible, et se sentent le moins mal possible » aurait la même valeur. Cela pourrait plutôt indiquer que l'atténuation des expressions émotionnelles négatives est moins valorisée pour le gain de valence que suivant un penchant aristotélicien pour le contrôle de soi. Les états émotionnels à valence négative tels qu'ils sont décrits ici (angoisse, colère) étant considérés comme moins délibérés et plus automatiques que le sourire (Ben-Ze'ev, 1997). Cette hypothèse explicative apparaît soutenue par le résultat suivant concernant les différences individuelles au MFQ.

Enfin, contrairement à d'autres études de psychologie morale dans lesquelles la dimension de pureté du MFQ corrèle négativement avec, par exemple, l'acceptabilité des dispositifs d'augmentation des capacités cognitives ou de mind-upload, l'acceptabilité ici était facilitée par l'adhésion des participants à cette dimension. Ceci peut suggérer que les transformations vocales ne sont pas perçues comme une atteinte à l'intégrité humaine, mais plutôt comme un moyen de gagner du contrôle sur soi. En d'autres termes ce filtre vocal serait perçu et apprécié comme un *outil anthropotechnique* d'auto-personnalisation (Goffette, 2006)). En cela, les motivations sous-tendant cette acceptabilité sembleraient rejoindre celles incitant à l'usage des filtres de beauté visuels pour lesquels il a été montré que le contrôle de l'image véhiculée était un déterminant considérable (Javornik et al., 2022). Pour autant, le potentiel anthropotechnique présenté par le filtre vocal apparaît ici comme dépassant la simple présentation de soi à autrui, la condition perception renvoyant plutôt un contrôle d'une part de ce que l'on « laisse entrer » dans son environnement et d'autre part à un gain de contrôle sur ses propres émotions en condition de vocal-feedback.

Limites :

L'une des limites évidentes de ce travail est l'accent mis sur un échantillon de participants occidentaux essentiellement jeunes et instruits (des étudiants), qui n'est

représentatif ni de la génération dans les pays occidentaux (comme le seraient, par exemple, des groupes d'enquête construits pour correspondre à la composition d'une population adulte donnée en termes de sexe, d'âge, d'éducation et d'appartenance ethnique - (Elias et al., 2019)), ni de la population plus globale des non-WEIRD (Henrich et al., 2010). Cela posé, ces résultats à la généralisation limitée correspondent néanmoins à un échantillon de population particulièrement sensible à ces nouvelles technologies au regard des travaux menés dans le cadre des filtres visuels (Javornik et al., 2022).

Par ailleurs, l'utilisation de vignettes dans les approches d'éthique expérimentale s'accompagne également de plusieurs limites. Tout d'abord, l'intensité des réactions suscitées par les histoires peut être limitée par la capacité d'immersion ou d'imagination du participant (Koverola et al., 2020b). Aussi, la lecture d'une vignette, en particulier celle décrivant une situation émotionnelle intense, peut ne pas susciter des réactions aussi fortes que dans les situations réelles correspondantes (Parkinson and Manstead, 1993). Dans le cas présent, nous avons tenté de remédier à ces limites en incluant une cover-story détaillée présentant le dispositif comme étant examiné en vue de sa commercialisation prochaine par une start-up réelle, et en précisant que les réponses des participants pèseront dans les décisions commerciales futures. Depuis juillet 2020, l'usage des technologies de transformation de voix a considérablement augmenté comme décrit en introduction, cela permet d'une part d'envisager une modification du jugement des participants, maintenant qu'ils sont plus familiers de cette technique et par ailleurs d'imaginer une réplication de ce type d'étude en remplaçant l'expérience de pensée demandée aux participants via les vignettes par une expérience in situ de ces filtres vocaux. Aussi, il convient également de noter que, même si nous avons conçu les 12 vignettes actuelles pour couvrir un large éventail de situations, la question reste ouverte de savoir si nos conclusions se généralisent à d'autres types de filtres vocaux et/ou à d'autres types de situations que celles testées ici.

Enfin, bien que ces résultats témoignent d'une grande acceptabilité quant aux filtres de modifications des émotions dans la voix, cette étude n'approfondit pas - au delà de nous renseigner sur l'influence des différences individuelles en termes de jugement moraux et de familiarité à la SF - les raisons sous-tendant la grande acceptabilité mesurée, déterminants qu'il serait essentiel de clarifier afin de guider la réflexion concernant la régulation de leurs applications.

2.5 Conclusion

A la lumière des résultats de notre étude, un filtre visant à atténuer la détresse émotionnelle dans la voix en général et plus spécifiquement chez un patient souffrant de TSPT tel que présenté dans une de nos vignettes-scénario (« *Imaginez que vous êtes une patiente souffrant de trouble de stress post-traumatique et vous consultez votre médecin car vous éprouvez de grandes difficultés à évoquer l'évènement traumatique qui vous est arrivé sans que votre voix ne se mette à trembler. Votre médecin vous propose d'utiliser les deux prothèses MyVoicePlus™ afin que vous (et vous seule) vous entendiez au quotidien avec une voix moins anxieuse, notamment dans des situations qui évoquent ce qui vous est arrivé. Ceci vous permettra de moins appréhender ces moments, et de mieux vivre avec votre maladie.* ») ne semble souffrir d'aucune objection morale dans l'échantillon de participants étudié. Au contraire, le contexte thérapeutique bénéficie d'une acceptabilité morale accrue, ainsi que les transformations visant à atténuer des émotions de valence négative.

Le chapitre suivant traite de l'importance de la voix dans le cadre spécifique de la thérapie d'exposition en imagination à destination des patients souffrant du TSPT. Au sein de deux études longitudinales menées dans cette population de patients, nous analysons les corrélats acoustiques de l'évolution de la symptomatologie psychotraumatique au cours des séances successives de psychothérapie, leur lien avec un autre marqueur physiologique (l'activité cardiaque) et discutons la portée de ces données, notamment sur le plan clinique.

3. Le Trouble de Stress Post Traumatique

Un cadre d'étude pour le filtre vocal

Comme nous l'avons exposé en introduction de ce manuscrit, cette thèse s'intéresse à l'utilisation de technologies informatiques capables de modifier certains paramètres de la voix et aux conséquences de celle-ci sur les processus cognitifs guidant nos inférences au sujet des états de notre interlocuteur en situation émotionnelle. De par l'extension du numérique au soin d'une part, et l'importance accordée à la voix et l'intensité émotionnelle caractérisant la situation de thérapie d'exposition en imagination pour le TSPT d'autre part, nous avons choisi de faire de cette situation le cadre de notre étude.

Ce chapitre développe la pertinence de la situation spécifique de la thérapie d'exposition en imagination pour étudier l'effet d'une situation émotionnelle particulièrement intense sur la voix des patients avant d'envisager dans un second temps, au Chapitre 4 les effets d'une transformation de cette voix sur les inférences perceptives des soignants. De fait, la présente situation thérapeutique constitue un matériel particulièrement riche qui offre notamment l'avantage d'être « naturel » par opposition aux corpus de voix d'acteurs ou encore aux protocoles expérimentaux d'induction émotionnelle qui sont traditionnellement utilisés dans le cadre des études sur les liens voix-émotions. A cette fin, nous consacrons une première partie de ce chapitre à détailler la symptomatologie du TSPT, les spécificités de sa prise en charge et l'usage possible du filtre de voix dans ce cadre. Dans une seconde partie, nous présentons le travail longitudinal réalisé chez deux cohortes de patients souffrant de TSPT afin de mettre en évidence les indices vocaux associés à cette situation.

3.1 Le Trouble de Stress Post Traumatique

Tout individu, bébé, enfant, adolescent, adulte, peut rencontrer au cours de sa vie des situations potentiellement traumatisantes ou traumatogènes. Aussi, la cinquième édition du Manuel Diagnostique et Statistique des Troubles Mentaux (DSM5) définit comme traumatogène toute situation qui implique ***une mort effective, une menace de mort, une blessure grave ou des violences sexuelles*** (American Psychiatric Association et al., 2013).

Les situations traumatogènes recouvrent les cas suivants (liste non exhaustive) :

— Violences volontaires physiques et sexuelles : maltraitance, abus, violences et exploitation sexuelles ;
— Accidents graves : accidents de la route, de transport, au travail, au domicile ;
— Maladie grave, soins palliatifs (de la personne elle-même ou d'un proche) ;
— Morts violentes par suicide, homicide, accident ou maladie ;
— Catastrophes naturelles, technologiques, accidentelles ;
— Attaques de masse, attentats ;
— Violences politiques et les guerres, les déplacements forcés ;
— Expositions traumatiques dans le cadre de l'exercice professionnel (soldats, policiers, pompiers, soignants, aidants, reporters de guerre, milieu carcéral, etc.) ;
— Situations de harcèlement institutionnelles, scolaires, professionnelles.

Chez un individu confronté à cette situation, cela peut entraîner l'apparition de troubles psychiques, dont le plus connu est le Trouble de Stress Post Traumatique (TSPT, également connu comme PTSD en anglais) qui pose de nombreux problèmes de santé publique de par son coût, sa chronicité possible et sa gravité. La prévalence vie entière de ce trouble en France au sein de la population générale est de 1 à 2 % de la population touchée (Vaiva et al., 2008).

3.1.1 Symptomatologie du TSPT

Un individu souffrant de TSPT présente un tableau clinique qui se subdivise en quatre catégories de symptômes : intrusions, évitement, altérations négatives de la cognition et de l'humeur, altérations de l'éveil et de la réactivité. Ces symptômes peuvent selon les cas, se développer à la suite directe de l'événement traumatique

(ET) ou apparaître à distance, parfois de plusieurs années après les faits. Ils sont au coeur du TSPT et essentiels à l'élaboration du diagnostique clinique.
En plus d'une anxiété traumatique spécifique, le patient peut se sentir coupable du fait de ses réactions au cours de l'ET ou parce qu'il a survécu alors que ce n'est pas le cas de tout le monde.

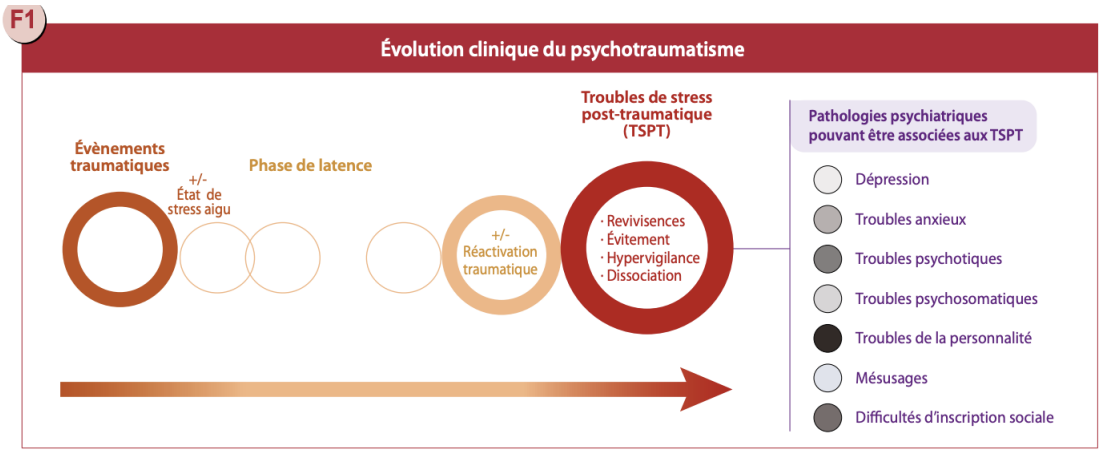


FIGURE 3.1 – Illustration de l'évolution clinique du TSPT tiré de l'article de Prieur et al. (2022)

Le diagnostic de TSPT repose sur plusieurs critères (American Psychiatric Association et al., 2013) : un individu doit avoir été confronté à la mort ou à une menace de mort, à une blessure grave ou à des violences sexuelles (Critère A), et présenter à la suite de cela des symptômes dans chacun des critères B, C, D et E décrits dans l'encadré suivant pour une période supérieure ou égale 1 mois (Critère F).

Principaux critères du trouble de stress post-traumatique (TSPT) selon le DSM-5 American Psychiatric Association et al. (2013)

Exposition à la mort, à une menace de mort, à une blessure grave ou à des violences sexuelles. Critère A

- exposition directe à l'évènement traumatique (ET)
- témoin direct (d'un ET survenu à d'autres personnes)
- annonce d'un évènement accidentel ou violent survenu à un proche (famille, ami)
- exposition répétée ou extrême à des situations aversives (par ex., soignants, secouristes, policiers)

Symptômes d'intrusion et reviviscence. Critère B, au moins un symptôme nécessaire :

- souvenirs répétitifs, involontaires et envahissants provoquant un sentiment de détresse,
- rêves répétitifs (par exemple, des cauchemars) provoquant un sentiment de détresse,
- réactions dissociatives : agir ou souffrir comme si l'événement se déroulait de nouveau, cela va des flash-backs à une totale perte de conscience de l'environnement présent,
- détresse psychique lors de l'exposition à des stimuli évoquant l'événement (par exemple, lors de la « date anniversaire » ou lorsqu'il entend des sons similaires à ceux entendus pendant l'événement),
- réactions physiologiques marquées lors de l'exposition à des stimuli évoquant l'évènement traumatique.

Symptômes d'évitement. Critère C, au moins un symptôme nécessaire :

- éviter les pensées, les sentiments ou souvenirs associés à l'événement,
- éviter les stimuli associés (par exemple, personnes, endroits, activités, situations, objets) à l'événement.

Altérations négatives des cognitions et de l'humeur. Critère D, au moins deux symptômes requis :

- perte de souvenir d'éléments importants de l'événement (amnésie dissociative),
- croyances ou attentes négatives tenaces et exagérées à propos de soi, des autres, ou sur le monde,
- pensées dysfonctionnelles persistantes sur la cause ou les conséquences du traumatisme qui conduisent à s'accuser soi-même ou les autres,
- état émotionnel négatif persistant (par exemple, peur, horreur, colère, culpabilité, honte),
- baisse marquée de l'intérêt ou de la participation à des activités pour lesquelles l'individu prenait autrefois plaisir
- sentiment de détachement ou d'éloignement des autres
- incapacité persistante à vivre des émotions positives (par exemple, le bonheur, la satisfaction, des sentiments affectueux).

Altérations marquées de l'éveil et de la réactivité. Critère E, au moins deux symptômes requis

- irritabilité ou accès de colère (avec peu ou pas de provocation),
- comportement imprudent ou autodestructeur,
- difficultés de concentration,
- réaction de sursaut exagérée,
- hypervigilance,
- perturbations du sommeil.

Source : DSM-5 : Diagnostic and Statistical Manual of Mental Disorders. Fifth edition

L'apparition de cette symptomatologie doit donner lieu à une perturbation qui entraîne une souffrance cliniquement significative ou une incapacité importante dans les dimensions sociale, professionnelle, ou toute autre dimension importante du fonc-

tionnement du patient (Critère G) et ne doit pas attribuable aux effets physiologiques d'une substance (par ex. médicament ou alcool) ou à une autre affection (Critère H).

3.1.2 Bases neurobiologiques des clusters de symptômes :

Les différents clusters de symptômes B, C, D et E ont des corrélats neurobiologiques, qui, bien que ne faisant pas directement l'objet d'investigation dans cette thèse, se doivent d'être abordés car ils sous-tendent le développement du TSPT et son évolution dans le temps, notamment en cours de thérapie, sujet qui nous occupe ici.

De plus en plus, les troubles psychiatriques - y compris le TSPT - sont considérés comme des troubles des réseaux cérébraux. En effet, des travaux récents suggèrent qu'un réseau cérébral élargi est impliqué dans les symptômes du TSPT. Les études de neuro-imagerie fonctionnelle ont permis d'identifier des régions cérébrales dont l'activité et la connectivité sont altérées. Le travail de Fenster et al. (2018) fait la revue du réseau cérébral impliqué dans chaque cluster de symptômes du TSPT (cluster du DSM-5) en prenant en considération les données issues des modèles neurobiologiques animaux et humains (voir figure 3.2 pour résumé). Les structures mises en évidence sont des régions particulièrement impliquées dans les processus de régulations émotionnelles,notamment des structures du lobe limbique comme l'amygdale, l'hippocampe et le cortex cingulaire, ainsi que du lobe frontal comme le cortex préfrontal médian et ventro-médian.

Sans les développer en plus de détails ici, ces données permettent de mieux comprendre les mécanismes impliqués dans l'apparition du trouble mais également dans la résilience, notamment lors de la thérapie. De fait, l'effet de la thérapie sur certains clusters de symptômes s'observe également par des modifications fonctionnelles au niveau des réseaux décrits précédemment. C'est le cas notamment des voix de signalisation de l'insula antérieure dont le rôle a été mis en évidence dans les phénomènes d'intrusion et qui voit son activité diminuer avec l'amélioration de ces symptômes en post-thérapie (Leroy et al., 2022).

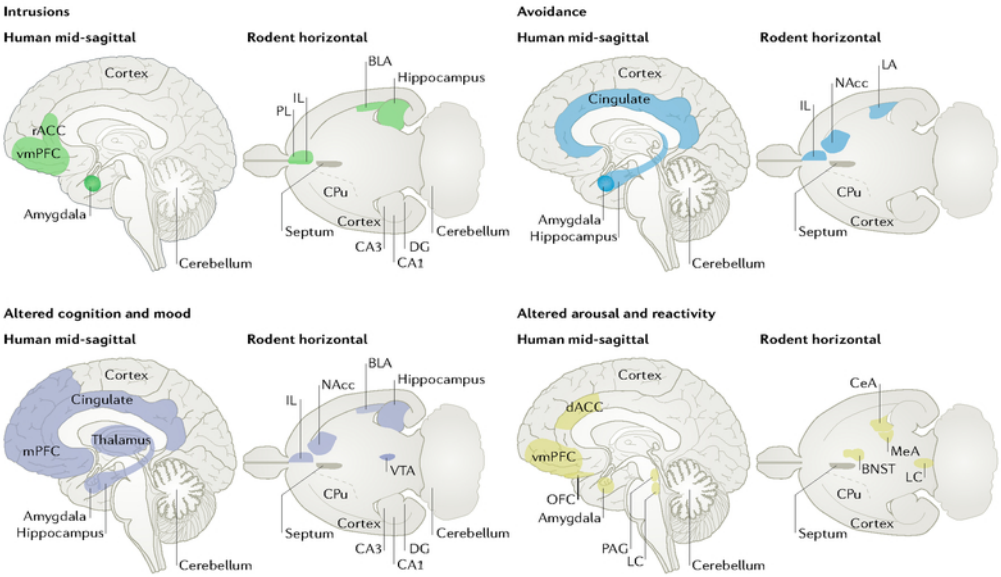


FIGURE 3.2 – Illustration des régions cérébrales impliquées dans le TSPT dans des études d'imagerie chez l'homme (coupe médiane sagittale), et des modèles murins (coupe horizontale). Chaque quadrant illustre les régions cérébrales pour lesquelles des données mettent en évidence un lien avec les symptômes du cluster.BLA, basolateral amygdala ; BNST, bed nucleus of the stria terminals ; CeA, central amygdala ; CPu, caudate and putamen ; dACC, dorsal anterior cingulate cortex ; DG, dentate gyrus ; IL, infralimbic cortex ; LA, lateral amygdala ; LC, locus coeruleus ; MeA, medial amygdala ; mPFC, medial prefrontal cortex ; NAcc, nucleus accumbens ; OFC, orbitofrontal cortex ; PAG, periaqueductal grey ; PL, prelimbic cortex ; rACC, rostral anterior cingulate cortex ; vmPFC, ventromedial prefrontal cortex ; VTA, ventral tegmental area. Figure adaptée de (Fenster et al., 2018)

3.1.3 Les répercussions du TSPT

Au delà des symptômes que nous avons décrits, les répercussions psychologiques et psycho-sociales du TSPT sont un sujet majeur et sont décrites dans une récente note de cadrage de l'HAS du 14 octobre 2020 dans les termes suivants :

> *« La souffrance psychique et l'affaiblissement associés aux syndromes psychotraumatiques peuvent être considérables et vont bien au-delà de ce que peut en dire un diagnostic psychiatrique. Au niveau psychologique, la littérature montre en effet qu'ils sont associés à de profonds bouleversements. Ces derniers peuvent se traduire par une remise en question parfois profonde du sens de la vie et de la logique du fonctionnement du monde. En outre, les syndromes psychotraumatiques altèrent le fonctionnement social, familial et*

professionnel des individus. Leur qualité de vie, ainsi que celle de leur entourage, s'en trouve largement diminuée avec des conséquences dévastatrices sur le plan individuel : chute de l'estime de soi, développement de sentiments de culpabilité et/ou de honte, stress, retrait social, etc. Sur le plan du pronostic fonctionnel et vital, l'impact des syndromes psychotraumatiques, et du TSPT en particulier, se révèle particulièrement délétère. En plus d'une majoration du risque suicidaire, les syndromes psychotraumatiques sont associés à de nombreuses comorbidités psychiatriques. Ainsi, l'étude de Breslau et al. de 1991 montre que près de 80% des participants souffrant de TSPT ont présenté ou présentent au moment de l'évaluation au moins un autre trouble psychiatrique, notamment un épisode dépressif majeur chez plus de 50% des sujets' (Breslau et al., 1991) » .

Les répercussions nombreuses du TSPT en font un enjeu de société notable.

Le TSPT un enjeu de santé publique en France

La prise en charge des personnes vivant l'épreuve du psychotraumatisme constitue un enjeu de santé public majeur. Un récent article de El-Hage et al. (2019) explicite les efforts récents menées en France quant à l'amélioration de la prise en charge des victimes de traumatisme.

« En France dans le contexte de menace terroriste décrit de ces dernières années et en réponse à la pression de l'opinion publique et des associations de victimes contre les violences domestique et sexuelle (les sondages suggéraient qu'il s'agissait de la principale préoccupation du public français), le président Macron a déclaré que la question de la violence à l'égard des femmes était une priorité de sa présidence. Le gouvernement a décidé de soutenir les centres de traumatologie publics dédiés à toutes les victimes, quel que soit le type de traumatisme ou les caractéristiques individuelles des victimes (âge, sexe, nationalité, etc.). Par la suite, en 2018, le gouvernement français a financé un centre national de ressources et de résilience (CN2R) et 10 services ambulatoires régionaux spécialisés en psychotraumatologie. Leur objectif est de garantir à long terme des soins de traumatologie de haute qualité à toutes les victimes qui en ont besoin sur l'ensemble du territoire » .

3.1.4 La thérapie d'exposition en imagination : assise théorique et principe

La thérapie cognitive émotionnelle et comportementale centrée sur le trauma par exposition prolongée (PE ; *prolonged exposure* en anglais) développée par Foa and Rothbaum (1998) est à ce jour la thérapie de première indication pour le TSPT (associée ou non à une prise de traitement médicamenteux ; McLean et al. (2022)). Le programme de la thérapie par exposition prolongée qui a montré son efficacité lors de nombreuses études (Foa and Kozak, 1986 ; Foa et al., 1991) est constitué de trois volets : une psycho-éducation sur les symptômes et la cause des difficultés associés au TSPT, une exposition in sensu ou en imagination, impliquant un récit répété du souvenir traumatique (reviviscence émotionnelle) et une exposition in vivo ou dans le monde réel aux souvenirs traumatiques, par exemple situations, objets lorsqu'ils sont sans danger.

Ici, nous nous focalisons en particulier sur l'exposition en imagination, qui constitue le contexte de nos travaux. Au cours de cette exposition en imagination à l'évènement traumatique (ET), le patient est amené à raconter à voix haute, à la première personne et au présent l'évènement traumatique qu'il a vécu et ceci au cours de plusieurs séances successives. Si la catharsis associée à l'évocation à voix haute de l'ET telle que décrite s'avère importante pour le processus d'acceptation et de guérison du patient, il ne s'agit pas là de l'objet du protocole. Celui-ci est étayé par un socle théorique qu'il est nécessaire d'expliciter. Ainsi, le terme « *prolonged exposure* » reflète l'appartenance de cette procédure à la longue tradition des thérapies d'exposition à destination des troubles anxieux, thérapies au cours desquelles les patients sont accompagnés dans la confrontation à des stimuli inoffensifs mais hautement anxiogènes afin de surmonter l'anxiété ou la peur excessive à laquelle ils sont associés (Brown et al., 2019). Sur le plan conceptuel, ce protocole thérapeutique repose donc sur la théorie du traitement émotionnel (*emotional processing theory* - EPT) qui articule des connaissances issues de modèles cognitifs et comportementalistes au sujet de l'apprentissage et la mémoire, le TSPT pouvant être observé comme « une pathologie du souvenir » (Brewin, 2018). Le point de départ de l'EPT est la théorie dite « bioinformationnelle » de l'imagerie émotionnelle de Lang (1977), qui décrit la peur comme une structure cognitive pensée comme un « programme » pour échapper au danger. Cette structure de peur inclut des

représentations au sujet du stimulus anxiogène (par exemple, le fameux ours menaçant de William James (Ellsworth, 1994)), des réponses associées à ce stimulus (*je m'enfuis*) et leur signification (*j'ai peur, les ours sont dangereux*). Lorsqu'une structure de peur renvoie à une peur réaliste, elle est considérée comme normale et sert alors de modèle d'action efficace à l'égard de menaces. Selon Foa and Kozak (1986) cette structure de peur devient pathologique dès lors que (1) les associations entre les stimuli ne sont pas représentatives du monde, (2) les réponses physiologiques et comportementales de fuite ou évitement sont déclenchées par des stimuli inoffensifs, (3) des réponses excessives interfèrent avec un comportement adapté et (4) des stimuli inoffensifs et les réponses sont associés de manière erroné avec la signification de peur.

Aussi, les théories neurocognitives du traumatisme psychique et du développement ultérieur d'un TSPT s'appuient largement sur les travaux ayant mis à jour la neurocircuiterie du conditionnement de peur. Le conditionnement de peur consiste à associer un stimulus neutre (CS) à un stimulus inconditionnel (UCS) qui déclenche intrinsèquement le comportement de peur (par exemple un choc électrique). En les présentant à plusieurs reprises de manière simultanée, UCS et CS vont donc être associés dans le cerveau. Après quelques essais, le CS génère à lui seul la réponse de peur, c'est la réponse conditionnée (CR). Le CS devient alors un prédicteur de l'expérience aversive initialement associée à UCS, ce qui va influencer le comportement (évitement, échappement...). chez l'animal et chez l'homme (Pitman, 1989). Les reviviscences (cluster B) et les symptômes d'hyper-éveil (cluster E) qui sont au centre du TSPT sont alors compris comme des réponses émotionnelles conditionnées qui résultent d'un conditionnement classique lors du traumatisme et qui sont ultérieurement déclenchées par les stimuli environnementaux. Les autres symptômes tels que les comportements d'évitement (cluster C), le détachement affectif (cluster D), etc... représenteraient alors dans cette perspective des tentatives pour gérer la détresse liée aux reviviscences du traumatisme.

Partant de ce modèle, Foa and Kozak (1986) ont proposé que les techniques d'exposition agissent en activant la structure de peur par l'exposition à des stimuli redoutés et en fournissant des informations rectificatives sur ces trois plans : les stimuli, les réponses et leur signification. Ainsi, un traitement émotionnel effectif nécessite deux conditions : l'activation de la structure de peur (sinon elle ne pourra pas être modifiée) suivie de la diminution de cette activation à la fois au cours de

la session d'exposition et entre les sessions d'exposition. La thérapie d'exposition en imagination repose donc sur le caractère labile de la trace mnésique et sur la théorie de la reconsolidation qui postule qu'un souvenir remémoré redevient instable avant de se consolider à nouveau en mémoire à long terme, d'où le terme « reconsolidation » (Brewin, 2018 ; Kindt, 2018 ; Sara, 2000). Cette phase de plasticité transitoire constitue alors une fenêtre d'action thérapeutique au cours de laquelle le souvenir peut être modulé, en particulier sa charge émotionnelle (Lonergan et al., 2013). C'est justement au sein de cette fenêtre de plasticité que nous envisageons l'usage d'un filtre vocal. Aussi, sur le plan clinique, une séance d'exposition en imagination à l'ET efficace nécessite donc que deux processus soient mis en oeuvre : la réactivation effective de la structure de peur associée à l'ET et la diminution de la réactivité émotionnelle au cours des séances successives. Un corollaire de cette exigence d'activation initiale est l'intensité de cette thérapie qui se doit d'engager le patient de manière particulièrement vive sur un plan émotionnel (Foa and Rothbaum, 1998). En effet des études ont montré que la mise en place de stratégies d'évitement et un faible engagement émotionnel pendant la thérapie d'exposition conduisaient à de moins bons résultats (Badour et al., 2012). Il s'agit de ce fait d'une thérapie éminemment difficile pour le patient et à cela viennent s'ajouter les symptômes d'évitement du cluster C qui rendent la tâche d'exposition à l'évènement particulièrement ardue. Ainsi, les taux d'abandons associés à cette thérapie varient de 17 % à 33 %, ceci pouvant être lié à la difficulté de la tâche demandée au patient(Bradley et al., 2005)). Dans ce cadre, « *tout ce qui peut atténuer la force émotionnelle du souvenir pendant la séance comporte un intérêt thérapeutique* » *(Brunet et al., 2011)*.

Le protocole d'exposition en imagination : un outil de recherche en neurosciences affectives

La méthode d'imagerie guidée par le script (« script-driven imagery SDI »), une forme proche de l'exposition pratiquée pendant la thérapie d'exposition en imagination est également utilisée comme protocole expérimental dans des études appelées « symptom provocation studies. » A cet égard, la SDI a été utilisée par les psychophysiologistes pour établir les corrélats biologiques du critère diagnostique B.5. du DSM-IV, « réactivité physiologique lors de l'exposition à des indices internes ou externes qui renvoient à un aspect de l'événement traumatisant » (American Psychiatric Association et al., 2013), et plus généralement à des fins de discrimination diagnostique (Orr and Roth, 2000). Elle a également été utilisée pour mettre en évidence les corrélats neuronaux de la réponse à l'exposition aux stimuli en lien avec l'évènement traumatique dans le TSPT (Lanius et al., 2006).

3.1.4.1 La thérapie d'exposition en imagination chimio-facilitée : une aide à la diminution de la charge affective

Parmi les traitements visant précisément ce processus de reconsolidation lié à l'exposition en imagination dans le TSPT, on trouve la thérapie chimio-facilitée par un médicament bêtabloquant, le propranolol. Même si les mécanismes sous-tendant cette facilitation sont encore débattus actuellement (Friedman, 2018 ; Kindt, 2018), trois essais cliniques ouverts, dont un avec des victimes non-remises de la catastrophe industrielle d'AZF (Toulouse), ont montré que la prise de ce traitement conjuguée à une séance d'exposition en imagination était associée à des diminutions significatives de symptômes post-traumatiques après six séances, améliorations qui se sont maintenues au suivi à 6 mois (Brunet et al., 2011).

Au cours d'un récent essai randomisé comparant une thérapie de réactivation chimio-facilitée par propanolol à une thérapie sous placebo (Brunet et al., 2018), les patients sous propranolol ont vu une diminution de leurs symptômes post-traumatiques de 36%, contre seulement 13% dans le groupe placebo. Bien que les résultats soient prometteurs, ils n'ont pas été suivis d'une évaluation à long terme. Une récente étude de Roullet et al. (2021) a répliqué ce protocole au sein de divers

centres en France et a mis en évidence une diminution significative de la symptomatologie après 6 séances de thérapie chimio-facilitée, un effet qui cette fois a été attesté comme durable 3 mois après la fin de la thérapie. Cette étude n'a cependant pas pu montrer la supériorité du groupe propanolol compte tenu de l'amélioration notablement importante relevée au sein du groupe contrôle (thérapie d'exposition sans propranolol), néanmoins les effets à 3 mois étaient quant à eux supérieurs au sein du groupe expérimental.

Pour autant, bien que le propanolol semble potentialiser l'efficacité de la thérapie de réactivation en imagination au souvenir traumatique, son utilisation reste soumise aux limitations habituelles des thérapies médicamenteuses. D'une part, elle exclut de nombreux patients présentant des contre-indications à la prise de béta-bloquants (contre-indication médicale à la prise de propranolol ou traitements en interaction avec le propranolol, prise d'un autre bêtabloquant ne pouvant être arrêté le temps du protocole, ECG significativement anormal, patients prenant des psychotropes pouvant entrer en interaction avec le propranolol). D'autre part, elle se heurte aux réticences de cette population de patients à suivre un traitement médicamenteux (Westenberg and Sandner, 2006). Dans ce cadre, une alternative non médicamenteuse permettant d'atténuer la force émotionnelle du souvenir serait d'un intérêt thérapeutique certain.

3.1.4.2 Le filtre vocal : une alternative non médicamenteuse visant l'atténuation de la charge émotionnelle

C'est dans ce contexte clinique que nous proposons de réfléchir à un filtre vocal ciblant les manifestations émotionnelles en jeu lors de la thérapie d'exposition en imagination.

La psychothérapie pour le TSPT est essentiellement un procédé qui vise à altérer la nature ou l'expression de ce qui a été appris pendant l'événement traumatique (Brewin, 2006). Un de mécanismes centraux pourrait ainsi être la recontextualisation du souvenir traumatique qui se produirait en invitant le patient à se remémorer l'évènement dans un contexte environnemental sécure (Brewin, 2018). Cela conduirait alors à la formation de nouveaux souvenirs contextualisés contenant l'information selon laquelle l'affect négatif causé par le rappel du souvenir est moins intense que prévu. Ce nouveau souvenir crée (le souvenir recontextualisé) pour lequel le trauma est associé à un contexte temporo-spatial plus sécure pourrait alors agir comme

une version du souvenir entrant en compétition avec la version originale ou alors remplacer le souvenir original - à ce sujet n'existe pas encore de consensus (Friedman, 2018). Quoiqu'il en soit, que ce soit en modifiant le souvenir original ou en créant un nouveau recontextualisé, il s'agira dans tous les cas de diminuer la charge affective de ce souvenir.

En de telles circonstances, nous considérons le potentiel thérapeutique d'un filtre vocal visant à diminuer les indices acoustiques de la charge affective lors de la thérapie. Un tel filtre vocal pourrait ainsi jouer un rôle comparable à celui du propranolol (Brunet et al., 2019 ; Roullet et al., 2021) et contribuer à une thérapie facilitée par le filtre de voix. Afin de caractériser les propriétés d'un tel filtre, il nous faut objectiver les marqueurs acoustiques de cette charge affective présents dans la voix de patients souffrant de TSPT pendant la thérapie d'exposition. Avant d'examiner les marqueurs acoustiques du TSPT, il convient de décrire les modalités de production de la voix ainsi que les possibilités de modifications synthétiques de celle-ci.

3.2 La voix

Note : Cette section est inspirée de l'article de Arias et al. (2021).

3.2.1 Principes de la production vocale.

La voix est produite quand un flux d'air, expiré par les poumons sous l'action des muscles thoraciques et abdominaux, met en vibration les cordes vocales (plus justement appelées les « plis » vocaux) situées dans le larynx. Cette vibration crée une onde acoustique qui se propage ensuite à travers le conduit vocal jusqu'à la bouche et au nez, puis qui rayonne dans l'environnement extérieur. Pendant son trajet dans le conduit vocal, la voix est filtrée par une multitude de parties anatomique mobiles (dites « articulateurs ») comme la langue et les lèvres dont la position, qui est sous contrôle volontaire du locuteur, amplifie certaines bandes d'énergie (ou « formants ») dans son contenu fréquentiel. Pour simplifier, la voix acquiert donc son intensité et sa hauteur (ou « fréquence fondamentale ») au niveau du larynx, puis son contenu spectral (ou timbre) au niveau des articulateurs (Titze and Martin, 1998a).

Si la mise en vibration des plis vocaux par l'air expiré est un processus passif, les propriétés mécaniques de ces plis (ex. leur tension), le flux aérien qui les met en mouvement, ainsi que les positions et propriétés de filtrage des articulateurs du conduit vocal sont tous contrôlables de façon active par plus de 100 muscles respiratoires, laryngés et oro-faciaux, innervés par des motoneurones dont les corps cellulaires sont situés dans la moelle épinière et le tronc cérébral. Ces motoneurones sont la voie de sortie d'une hiérarchie complexe de systèmes neuronaux corticaux et sous-corticaux, qui impliquent notamment le cortex moteur laryngé (laryngeal motor cortex, LMC), le cortex cingulaire antérieur (ACC) et, dans le mésencéphale (midbrain), la substance grise périaqueducale (PAG) (Simonyan and Horwitz, 2011).

Ces influences corticales et sous-corticales sur les actuateurs musculaires situés à toutes les étapes de la chaîne de production vocale (expiration, phonation et articulation) ont une multitudes d'effets complémentaires, additifs ou soustractifs, sur le contenu acoustique final de la parole. Ces effets acoustiques permettent d'implémenter une multitude de fonctions langagières, généralement regroupées sous le terme de « phonétique » : la qualité de voix (par exemple, une voix claire ou rauque), le façonnement spectral des segments individuels (par exemple, la couleur spectrale particulière permettant de distinguer un /a/ d'un /o/, ou le motif temporel d'occlusion de la bouche permettant de distinguer un /t/ d'un /b/), et les traits supra-segmentaux de la parole (par exemple, les accents toniques, l'intonation ou le rythme) - ces derniers étant souvent appelés « prosodiques » (Dubois et al., 1974).

Au niveau des muscles thoraciques qui sont contrôlés depuis la corne ventrale de la substance grise de la moelle épinière, les changements de la pression d'air sous-glottale se traduisent principalement par des changements de l'intensité vocale. Ces modulations d'intensité peuvent par exemple communiquer un état émotionnel plus intense (ex. voix excitée et joyeuse plutôt que calme et triste) (Ilie and Thompson, 2006) ou, de façon plus locale, marquer l'emphase sur un mot particulier (« c'est *Marie* qui a mangé le gâteau »). A des niveaux de pression sous-glottale plus forts, comme dans les cris de douleur ou de colère, mais aussi possiblement dans le cas de contrôle neurologique altérés sur les muscles du larynx, comme dans le stress ou l'anxiété, la vibration des plis vocaux peut également quitter son régime de fonctionnement habituel (dit « linéaire » ou « harmonique ») et produire des oscillations irrégulières (dites « non-linéaires » ou « inharmoniques ») qui ont des conséquences audibles sur

le timbre de la voix : voix rauque, rugueuse, bruitée ou soufflée. Ces modulations, volontaires ou involontaires, de la qualité vocale ont une importance particulière pour la communication des émotions, et sont associés à certaines mesures que nous utiliserons dans l'étude 3.4 et seront décrits en section 3.4.1.6.

Au niveau des muscles laryngés, qui sont innervés par le nerf vague (X) également impliqué dans la régulation végétative (digestion, fréquence cardiaque, etc.), les changements de longueur, d'ouverture et de rigidité des plis vocaux changent leurs propriétés vibratoires au passage du flux d'air en provenance des poumons. Ces changements ont des conséquences sur la qualité de voix (comme mentionné plus haut), mais surtout sur la fréquence fondamentale de la voix (ou f_0), qui gouverne la sensation de hauteur (ou *pitch* en anglais). Une équation approximative reliant la fréquence de vibration f_0 aux propriétés des plis est :

$$f_0 = \frac{1}{2L}\sqrt{\frac{s}{d}} \tag{3.1}$$

où L est la longueur des plis vocaux (corrélés à la taille de l'individu), s est la tension exercée par les muscles sur les plis, et d est la densité des tissus constituant les plis (Titze and Martin, 1998b). Le pitch de la voix est donc inversement proportionnel à la longueur des plis et directement proportionnel à la racine carrée de la tension sur les plis : les plis plus longs et moins tendus produisent une voix plus grave. Entre individus, ces propriétés mécaniques liées à la taille expliquent notamment le dimorphisme sexuel entre les voix d'hommes et de femmes, dont le pitch pour ces dernières est en moyenne 2 fois plus élevé (Puts et al., 2006). Au niveau individuel (« à taille constante »), une voix en moyenne plus aiguë ou plus grave peut correspondre à des états émotionnels différents (Ilie and Thompson, 2006) ; des variations de f_0 plus exagérées peuvent communiquer des états plus activés, comme la peur par opposition à la tristesse (Pell and Skorup, 2008) ; enfin, une augmentation de f_0 au début ou à la fin d'une phrase peut signifier certaines modalités linguistiques (ex. question ou affirmation) ou certaines attitudes (ex. doute, ironie) (Jiang and Pell, 2017).

Les muscles oro-faciaux, qui sont innervés par les nerfs faciaux (VII) et trijumeaux (V) prenant leur origine dans le pont (pons), contrôlent les articulateurs de la région supra-glottale (ex. les lèvres, la langue, la mâchoire) et ont la capacité de modifier de façon très rapide la forme et les propriétés résonantes du conduit vocal (voir

par ex. (Frahm, 2018) [video] Beatboxing in real time. Available from : https://www.youtube.com/watch?v=Wh4aEc4yPh0). Acoustiquement, ces changements se traduisent par une amplification ou diminution de l'énergie dans certaines bandes de fréquences (ou « formants ») du signal vocal, avec la conséquence de modifier non pas la f_0 (qui est dictée par la fréquence de la vibration des plis vocaux) mais le *timbre* de la voix (d'un point de vue psychoacoustique, le timbre est défini comme « cet attribut de la sensation auditive qui permet à un auditeur de distinguer 2 sons non-identiques ayant la même intensité et le même pitch » - (ASA, 1994)). Ces changements de timbre sont mobilisés pour distinguer différentes voyelles ou consonnes : par exemple, l'articulation d'un son /a/ en français se distingue d'un /o/ par une position de la langue (lieu d'articulation) plus antérieure et un plus grand écartement de la langue au palais (aperture). Ils sont également audibles dans plusieurs gestes faciaux émotionnels, comme l'expression de dégoût qui diminue la conduction aérienne par le nez (Chong et al., 2018) , ou le sourire, qui raccourcit la longueur du conduit vocal en étirant les coins de la bouche par les muscles zygomatiques (Arias et al., 2018b).

3.2.2 Techniques informatiques de transformation de la voix

Plusieurs techniques informatiques permettent aujourd'hui de simuler ces changements acoustiques sur une voix déjà enregistrée, et d'en changer ainsi le sexe perçu, l'identité, mais aussi de façon plus dynamique, les émotions ou les attitudes, et ce parfois en temps-réel (Arias et al., 2021).

Pour manipuler la fréquence fondamentale (f_0) d'un enregistrement vocal, la technique la plus simple consiste à modifier la vitesse (ou le « taux d'échantillonnage ») avec laquelle cet enregistrement est lu : s'il est lu plus rapidement, les maxima d'amplitude de l'oscillation vocale se rapprochent, et donnent l'impression d'une fréquence plus élevée et d'un pitch plus aigu. Cette méthode, dite de « ré-échantillonnage » ou de « lignes de retard multiples » (Dattorro, 1997) présente l'intérêt de pouvoir réaliser cette transformation avec un temps de latence très court (de l'ordre de 20 millisecondes - (Rachman et al., 2017)), mais introduit des artefacts de timbre qui deviennent rapidement audibles au delà de changements de hauteur très faibles (de l'ordre du demi-ton). En échantillonnant le signal à une vitesse différente, on modifie en effet non seulement la f_0 de la source glottale, mais on déplace également les

fréquences de résonance du conduit vocal : on ne peut pas ainsi augmenter la hauteur d'une voix d'homme sans modifier, dans un premier temps, le timbre des voyelles qu'ils prononcent (ex. les /o/ commencent à sonner comme des /a/) voire même son identité : sa voix commence à sonner comme celle d'une femme, d'un enfant, voire même d'un petit animal (effet souvent décrit comme « chipmunk » ou « mickey mouse »).

Pour éviter ces artefacts, les méthodes plus modernes de modification de hauteur s'efforcent de séparer les informations de source glottale (à transformer pour changer la f_0) et celles du conduit vocal (qui sont à préserver pour garder un timbre constant malgré le changement de hauteur). Ces méthodes travaillent typiquement à partir d'une analyse spectrale du signal par transformée de Fourier locale (*short-time Fourier transform*), à partir de laquelle on analyse le spectre du signal toutes les 50-100ms, on repère les fréquences fondamentales et leurs harmoniques, et on effectue un certain nombre de transformation pour décaler celles-ci de façon cohérente avant de reconstruire le signal d'origine par transformée de Fourier inverse. Ces méthodes incluent par exemple la technique PSOLA (pitch synchronous overlap and add) implémentée dans le logiciel PRAAT (Boersma, 2001) ; le vocodeur de phase (Moulines and Laroche, 1995) implémenté dans le logiciel CLEESE (Burred et al., 2019) ; ou la méthode d'interpolation adaptative STRAIGHT (Kawahara, 1997). Même si ces méthodes font encore aujourd'hui l'objet de recherches et d'optimisation intenses, notamment dans l'industrie de la production musicale où elles sont beaucoup utilisées, elles permettent généralement d'obtenir des transformations de bonne qualité pour des manipulations de hauteur d'amplitude raisonnables (inférieures à une octave). Cependant, comme elles nécessitent une analyse du signal sur des fenêtres de 50-100ms, leur capacité de temps-réel est plus limitée que les méthodes à base de ré-échantillonnage : si une latence de 100ms, comparable par exemple à celle introduite par des outils de visioconférence comme Zoom ou Skype, est tolérable pour une conversation, elle devient en effet trop grande pour fournir à un locuteur un retour vocal temps-réel de sa propre parole (on commence à bégayer autour de 50ms de latence ; (Stuart et al., 2002). Il faut cependant noter que, comme partout, les progrès récents de l'intelligence artificielle permettent d'introduire de nouvelles méthodologies de transformation qui pourraient rapidement résoudre ce compromis latence/qualité : en particulier, certaines implémentations de vocodeur de phase par réseaux de neurones

profonds permettent maintenant de conditionner l'apprentissage de transformations temporelles (de type wavenet) par un paramètre de hauteur modifiable (Wu et al., 2019).

3.3 Voix et TSPT

Comme nous l'avons vu, la production de la voix et sa qualité qui en résulte sont le fruit d'une intrication de facteurs anatomiques et psychophysiologiques, les états émotionnels relevant des ces derniers. Dans ce cadre, la voix porterait les stigmates de la dérégulation émotionnelle (Gross and Jazaieri, 2014), cette dernière caractérisant nombre de désordres psychiatriques dont le TSPT.

A cet égard, quelques travaux se sont penchés sur la question de la voix dans le TSPT, avec principalement pour objectif de mettre en évidence des biomarqueurs acoustiques de la pathologie. Compte tenu de la nature de leur objectif, l'approche adoptée consiste alors à utiliser l'analyse automatique via l'apprentissage machine afin d'extraire des régularités dans de larges corpus de sons (Scherer et al., 2015) et de créer un outil de dépistage vocal systématisé (Leightley et al., 2019 ; Marmar et al., 2019 ; Xu et al., 2012). Un des arguments associé à cette entreprise est celui de la rentabilité, comme explicité ici par Xu et al. (2012) : *« Il est souvent difficile et coûteux de diagnostiquer le TSPT en raison de la difficulté d'accès des patients et de la variabilité des symptômes présentés. Un système de dépistage vocal automatisé capable de repérer à distance les personnes présentant un risque élevé de TSPT et de suivre leurs symptômes pendant le traitement pourrait contribuer à lever les obstacles à l'évaluation rentable du TSPT[...] et peut être utilisé pour un suivi continu et à distance du TSPT sans nécessiter de visites coûteuses et fréquentes d'un clinicien »*.

Plusieurs études se sont attelées à créer un outil de dépistage automatique. Ainsi Xu et al. (2012) rapportent que leur système appelé Tele-PTSD-monitor conçu à partir d'un corpus de voix issues de personnes souffrant ou non de TSPT, est capable de détecter le TSPT avec une précision de 95.88% à partir de quelques secondes de données seulement. Ces résultats bien qu'encourageants souffrent néanmoins de certains biais méthodologiques. Tout d'abord, l'établissement du diagnostique de TSPT ne repose pas sur un clinicien mais uniquement sur un autoquestionnaire et le matériel sur lequel a été réalisé l'analyse est très peu contrôlé car constitué de

10 vidéos disponibles au public (5 de personnes souffrant de PTSD, 5 non PTSD). Une autre étude plus récente de Marmar et al. (2019), évite quant à elle ces biais méthodologiques en analysant une large cohorte d'extraits vocaux provenant de 52 malades et 77 contrôles dont le TSPT était évalué via la CAPS (Clinician Administered PTSD Scale CAPS-IV - (Nader et al., 1996)), un entretien diagnostique standardisé et validé, et rend compte d'un taux global de classification correcte de 89,1%. Néanmoins, une étude récente (Leightley et al., 2019) évaluant différents classifieurs reposants sur une méthode d'apprentissage automatique supervisé pour l'identification de TSPT chez d'anciens militaires au Royaume-Uni a montré que si une sensibilité satisfaisante est obtenue par certains classifieurs, la spécificité reste faible, se pose donc le risque de diagnostics faussement négatifs. Outre, les possibles biais méthodologiques, il est nécessaire de noter que ces travaux utilisent des corpus de parole quotidienne (par exemple collecté par smartphone) et non pas de la voix de patients en cours de thérapie d'exposition en imagination. Cette différence nous permet ainsi de pointer la spécificité de notre question qui est d'examiner les marqueurs de la charge affective associés au TSPT précisément pendant le récit personnel que le patient fait de l'ET qu'il a vécu, une situation qui n'équivaut pas à poser la question d'une « voix du TSPT ». En outre, l'accent qui a été mis sur la discrimination diagnostique dans le TSPT et plus largement en psychiatrie tend nouvellement à être critiqué car il ne permettrait pas de rendre compte de l'hétérogénéité de l'expérience subjective inhérente à chaque parcours (Taschereau-Dumouchel et al., 2022).

D'autres études, plus proches de notre contexte de thérapie ont été menées hors cadre de l'apprentissage machine. Ainsi, Scherer et al. (2013) montrent qu'en situations émotionnellement positives, négatives et neutres créées expérimentalement, les personnes souffrant de TSPT présentaient des caractéristiques vocales plus tendues ainsi qu'une contraction de l'espace formantique (Scherer et al., 2015). Dans la même veine, van den Broek et al. (2011) ont demandé à des personnes souffrant de TSPT de produire deux récits positif et négatif et mettent en avant 65 paramètres acoustiques expliquant 69 à 83% de la variance des symptômes de stress chez ces patients. Bien que plus proches de nos conditions d'étude, étonnamment les situations émotionnelles négatives des deux expériences ne relevaient pas de l'ET des patients. De plus, l'absence de mesures répétées dans le temps ne permet pas d'une part de s'affranchir des différences individuelles de la voix, les paramètres acoustiques mis en évidence

pourraient relever de manière indéterminée du TSPT mais également de spécificités autres de la cohorte examinée, d'autre part elle ne permet pas de saisir l'évolution de la voix en relation avec celle des symptômes qui ont une dynamique propre au delà de la distinction binaire malade/non malade.

A ce sujet, O'Donnell et al. (2007) soulignent l'importance d'adopter ici une approche longitudinale. « *L'identification des trajectoires de symptômes du TSPT au cours du temps a d'importantes implications théoriques et pratiques, avec notamment le potentiel d'informer notre compréhension de la phénoménologie du TSPT et du processus de guérison* ». Pourtant, à notre connaissance, aucune étude sur la voix dans le TSPT ne propose de cohorte longitudinale permettant ainsi d'adresser la question cruciale des processus à l'oeuvre dans la pathologie et sa rémission.

3.4 Etude TraumacoustiK

Étude longitudinale des corrélats acoustiques du TSPT dans la voix pendant la thérapie d'exposition en imagination à l'évènement traumatique

Nous proposons donc ici de constituer une cohorte longitudinale de patients permettant d'analyser l'évolution de la voix de patients souffrant de TSPT tout au long de leur prise en charge en thérapie d'exposition en imagination à l'ET. Cette étude clinique est officiellement enregistrée sous l'appellation « TRAUMACOUSTIK - Recherche des corrélats bio-acoustiques de la symptomatologie du TSPT et de son évolution au cours de la thérapie d'exposition en imagination à l'évènement traumatique ».

Pour ce faire, nous choisissons, malgré le mouvement actuel dans le domaine de l'analyse de la voix consistant à extraire un maximum de paramètres acoustiques via la puissance des analyses d'apprentissage machine, d'utiliser une méthode d'analyse statistique consistant plutôt à extraire un nombre limité de paramètres acoustiques (décrit en section 3.4.1.6, choisis pour être classiquement associés aux états émotionnels (Juslin and Västfjäll, 2008 ; Scherer, 2003 ; Scherer et al., 2001). Nous utilisons ensuite des modèles de régression pour mettre en évidence des régularités reliant la symptomatologie du TSPT à ces paramètres acoustiques, qui permettent une interprétation phonétique plus claire (« je sais dire ce qu'est une augmentation de pitch, je ne

sais pas dire ce qu'est une augmentation de MFCC » - dirait le traiteur de signaux). Ceci nous paraît constituer des conditions plus propices à « faire naître du sens clinique » à partir de ce travail et à la compréhension des mécanismes de la pathologie. Notre étude se concentrera alors sur le pitch ainsi que sur des indices témoignant de la qualité vocale à savoir le jitter (quantifiant les fluctuations de la période fondamentale du signal d'un cycle à l'autre), le shimmer (miroitement, variations locales d'amplitude) et la proportion de bruit dans le signal (noise-to-harmonic ratio). Enfin, en faisant le choix d'analyser la voix des patients pendant qu'ils décrivent l'ET qu'ils ont vécu, nous nous inscrivons dans une démarche plus générale s'attachant à mettre l'accent sur la spécificité et l'hétérogénéité des parcours individuels dans la recherche sur les troubles anxieux. En cela, plusieurs chercheurs de la communauté des neurosciences dont Joseph LeDoux, chercheur reconnnu pour ses travaux sur le rôle de l'amygdale dans la peur, dans un récent article au titre évocateur « *Putting the "mental" back in "mental disorders" : a perspective from research on fear and anxiety* », propose qu'un facteur, plus que les autres a contribué au manque d'efficacité des traitements des troubles anxieux à savoir « la marginalisation systématique de l'expérience subjective du patient en tant que sujet de recherche et cible de traitement » (Taschereau-Dumouchel et al., 2022).

3.4.1 Matériel et méthodes

3.4.1.1 Encadrement réglementaire de l'étude : aspects éthique et légaux

Conformément au Code de la Santé Publique, le promoteur ici Fédération régionale de recherche en psychiatrie et santé mentale Hauts-de-France (F2RSM Psy) a soumis une demande d'avis auprès du CPP avant le début de la recherche et adresse une copie et un résumé de la recherche à l'ANSM une fois l'avis favorable du CPP obtenu. Le projet « TRAUMACOUSTIK - Recherche des corrélats bio-acoustiques de la symptomatologie du TSPT et de son évolution au cours de la thérapie d'exposition en imagination à l'évènement traumatique » a reçu l'autorisation du Comité de Protection des Personnes Île de France 1 (CPPIDF1-2019-ND62-cat.3) NSI19.06.27.49144 en otcobre 2019. La mise en place officielle du protocole a eu lieu le 20 janvier 2020.
Il s'agit d'une étude non interventionnelle, observationnelle qui ne modifie pas la pratique courante, sans procédures invasives supplémentaires ou inhabituelles de diagnostic ou de surveillance. Elle répond donc aux critères d'une étude RIPH de

Catégorie 3. Chaque patient éligible s'est donc vu remettre une note d'information afin de collecter leur non-opposition quant à la participation à cette étude.

3.4.1.2 Financement

Étude financée par l'AAP Émergence 2019 du CHRU de Lille.

3.4.1.3 Participants

La population de notre étude est constituée de 20 patients (12 femmes) reçus au Centre Régional Psychotrauma Hauts-de-France (CHRU de Lille) (dirigé par le Dr Frédérique Warembourg).
Cet effectif a été jugé suffisant pour mettre en évidence des corrélations comme cela a pu être le cas dans l'étude de Ellgring and Scherer (1996) parmi leur échantillon de N=11 patientes souffrant de dépression (corrélations significatives entre paramètres acoustiques et scores à l'échelle de dépression en mesures répétées).

Le Centre Régional Psychotraumas Hauts-de-France

Ce centre régional est dédié au diagnostic et à la prise en charge des patients souffrant d'un TSPT. Sur le plan clinique, il est composé de 9 psychiatres (3 ETP), de 7 psychologues (2,5 ETP) et de 2 internes en psychiatrie, formés à la thérapie par exposition prolongée, d'une infirmière pour réaliser les évaluations de premières demandes conjointement avec les médecins-internes. Les patients y sont adressés soit par leur psychiatre, leur centre médico-psychologique ou leur médecin traitant, soit sont orientés par les urgences (urgences psychiatriques ou générales), l'unité médico-judiciaire du CHU de Lille, ou des associations de victimes, ou prennent RDV eux-même (coordonnées du Centre Régional Psychotrauma accessible sur internet).

Critères d'inclusion :
Les patients sont recrutés suivant les critères ci-après.
— âgé d'au moins 18 ans
— réponse aux critères DSM-5 de Trouble de Stress Post-Traumatique
— PCL-5 strictement supérieure à 32 (PCL-5, Weathers et al., 2013)

— formulation de la non-opposition
— patient dont le français est la langue maternelle
— traitement stable depuis au moins 2 mois et pendant toute la durée de l'étude

Critères de non-inclusion :
— patient sous sauvegarde de justice, sous tutelle ou sous curatelle.
— trouble psychotique
— patient sourd, muet ou ayant un trouble du langage pouvant affecter la prosodie
— risque suicidaire sévère avéré (évalué au moyen du Mini-International Neuropsychiatric Interview-Suicidality module ; MINI-S)
— toxicomanie aux opiacés ou dépendance alcoolique actuelles
— patient non affilié à un régime de sécurité sociale
— femme enceinte ou allaitante.

Pour notre étude, nous avons choisi de recruter indistinctement des patients souffrant de TSPT simple consécutif à un évènement isolé (accident, attentats, agression, etc.) ou de TPST complexe consécutif à l'exposition répétée à des violences (maltraitance, violences intrafamiliales, professions à risques, etc.).

Modalité d'identification des participants :
Dans le cadre de cette recherche, les sujets sont identifiés de la façon suivante : n° ordre de sélection de la personne dans le centre (2 positions numériques) - initiales nom et prénom. Cette référence est unique et est conservée pour toute la durée de la recherche. Elle est utilisée pour les documents papiers (Case Report Form et auto-questionnaires) ainsi que pour les fichiers audio.

3.4.1.4 Procédure

Les patients sont rencontrés au CRP HdF (Hôpital Fontan - CHRU de Lille). Le premier entretien pose ou confirme le diagnostic de TSPT. Ce diagnostic d'entrée dans l'étude est évalué par le clinicien et déterminé par le module TSPT de l'entretien diagnostique structuré du Mini-International Neuropsychiatric Interview (MINI ; voir section Mesures cliniques), selon le mode de prise en charge clinique habituel. Le principe de la thérapie d'exposition par imagination à l'évènement traumatique est ensuite expliqué au patient. C'est alors que notre étude leur est proposée. En cas d'accord de participation, nous procédons au recueil de la non opposition.

S'ensuivent alors les différentes visites inhérentes au protocole que nous décrivons ici :

1. **Visite d'inclusion :** La visite d'inclusion (V0) nous permet de vérifier l'éligibilité du patient quant à sa participation à notre étude selon les critères d'inclusion et non-inclusion explicités plus haut. La comorbidité psychiatrique (actuelle et passée) constituant les critères d'exclusion est évaluée par le clinicien via le MINI. Cette visite permet également le recueil des informations démographiques (âge et sexe du patient) et informations cliniques (recueil des traitements) nécessaires à l'étude.
2. **Visite 1 : Première séance de thérapie d'exposition en imagination :** La première visite (V1) se déroule conformément à la prise en charge habituelle dans le cadre de ce protocole thérapeutique. Le clinicien débute la visite par un entretien clinique et fait remplir au patient la PTSD Checklist for DSM V (PCL5, voir section 3.4.1.5) : auto-questionnaire visant à évaluer les symptômes du TSPT ressentis depuis la dernière visite V0. S'en suit l'écriture détaillée du script de l'événement traumatique par le patient encadrée par le thérapeute. Le script est ensuite lu à voix haute par le patient. Cette première lecture fait l'objet d'un enregistrement de la voix du patient.
3. **Visites de suivi V2 à V5 (ou plus) : Déroulement de thérapie d'exposition en imagination :** Chaque visite de la thérapie (hebdomadaire ou bimensuelle) respecte ensuite le modèle suivant :
   — Entretien clinique
   — Évaluation des symptômes depuis la dernière visite via la PCL-5
   — Lecture du script par le patient.
   — Enregistrement de la lecture du script selon la même procédure qu'à la visite V1.

Le nombre de séances requis pour chaque patient est décidé par le thérapeute, en fonction de l'évolution de la symptomatologie. Celle-ci est évaluée par un entretien clinique et aidée du score à l'autoquestionnaire PCL-5. Le nombre de séances d'exposition minimum est de 5 séances. Ce chiffre peut se voir augmenter si le score PCL-5 reste strictement supérieur au seuil de 32.

3.4.1.5 Mesures cliniques

Le Mini-International Neuropsychiatric Interview (MINI-DSM-V) (Crocq et al., 2016) est un entretien diagnostique structuré, d'une durée de passation brève , explorant de façon standardisée, les principaux troubles psychiatriques répertoriés au sein du DSM-V.

La PTSD Checklist - Version 5 (PCL-5, (Weathers et al., 1993) est une échelle de 20 items qui évalue les critères diagnostiques du TSPT. La quantification de l'intensité des symptômes du TSPT proposée par la PCL est d'une grande prédictibilité pour un diagnostic de TSPT. De très nombreuses études psychométriques ont montré la validité et la fidélité de la PCL dans de nombreux échantillons (vétérans, étudiants, accidentés, victimes de violences etc.) et attestent de sa sensibilité et de sa spécificité pour un diagnostic de TSPT, de sa cohérence interne et de sa fidélité très satisfaisantes, ainsi que de sa validité convergente et discriminante. Il faut ajouter que la PCL est parfaitement adaptée à une utilisation dans de larges échantillons. Le participant doit rapporter pour chacun des 20 symptômes de TSPT l'intensité ressentie selon une échelle de modalités ordonnées (0 = pas du tout à 4= très souvent) ; ainsi le score total peut aller de 0 à 80. La version francophone de la PCL a été validée et montre qu'un score total > 33 est corrélé à un diagnostic de TSPT selon le DSM 5.

3.4.1.6 Mesures acoustiques

La phase de lecture à voix haute du script de traumatique de chaque visite est enregistrée via un appareil enregistreur Zoom H4n Pro avec les microphones stéréo positionnés en XY (90/120°) à 30 cm du patient. L'enregistrement est fait « en qualité CD », avec un taux d'échantillonnage de 44.1 kHz et une quantification sur 16bits.

Chaque enregistrement est dans en premier temps écouté et nettoyé manuellement sur Audacity (© 1999-2021 Audacity Team) en respectant la procédure suivante :

— les silences et interventions du thérapeute en début et fin d'enregistrements sont coupés
— l'amplitude du son de chaque fichier est normalisée

— les périodes de prise de parole du thérapeute en cours de script et les bruits divers (porte, tourne de page, grincements...) sont remplacées par des périodes de silence de durée identique à la période remplacée.

Une fois les fichiers sons nettoyés, une analyse acoustique automatisée est réalisée sur chaque enregistrement avec le logiciel PRAAT (Boersma, 2001). Les caractéristiques acoustiques implémentées dans PRAAT (et leur définition synthétique) que nous avons extraites pour chaque enregistrement sont :

— le *mean pitch* qui correspond à la moyenne des valeurs de pitch extraites toutes les secondes,
— le *jitter local*. Il s'agit de la différence absolue moyenne entre des périodes consécutives, divisée par la période moyenne,
— le *shimmer apq3*. Il s'agit du quotient de perturbation de l'amplitude en trois points, c'est-à-dire la différence absolue moyenne entre l'amplitude d'une période et la moyenne des amplitudes de ses voisines, divisée par l'amplitude moyenne,
— le *mean nhr*, rapport bruit sur harmonique moyen sur l'extrait.

Afin de fluidifier la lecture du manuscrit nous utiliserons dans la suite de cette thèse, les termes pitch, jitter, shimmer et nhr pour renvoyer aux mesures précises que nous venons de décrire.

La table 3.1 présente les valeurs moyennes des différents paramètres et leurs écarts types pour toutes les séances chez les patients de notre cohorte séparés en hommes et femmes.

sex	mean pitch	std pitch	mean jitter	std jitter	mean shimmer	std shimmer	mean hnr	std hnr
femmes	177.63	33.67	3.14	0.49	17.02	3.39	10.2	2.15
hommes	115.78	15.49	3.22	0.49	17.97	3.23	8.85	1.51

TABLE 3.1 – Résumé des valeurs moyennes et écarts types pour les paramètres acoustiques extraits dans l'étude TraumacoustiK

3.4.1.7 Analyses statistiques

Les analyses statistiques que nous présentons sont faites en Python, avec les modules `seaborn` (© M. Waskom, 2021-2023) `pingouin` (© R. Vallat, 2018-2023) et `pymer` (© E. Jolly, 2017-2023). Pour les analyses de régression, `pymer` utilise le module

`lmer` qui estime le modèle testé à l'aide de la méthode de vraisemblance restreinte ou résiduelle (acronyme anglais REML) et les p values ainsi que les degrés de libertés (DF) sont estimés par l'approximation de Satterthwaite.
Les données ont été analysées en utilisant des modèles de régression linéaire mixtes (*linear mixed models*, LMM) pour lesquels nous présentons l'équation du modèle (dans sa syntaxe *R* (Banse and Scherer, 1996)). La pertinence de chaque paramètre du modèle sera illustrée par les résultats du test T de Student de nullité statistique du coefficient associé à chaque paramètre (H0 : $\beta = 0$, contre H1 : $\beta \neq 0$) sous la forme de la valeur du coefficient β d'estimation du paramètre, la valeur du t et sa p-value.
Les analyses suivantes sont réalisées suivant un plan expérimental intra-patient.

3.4.2 Résultats

Les données que nous présentons ont été collectées entre Janvier 2020 et Décembre 2021 (avec des périodes de coupures liées aux restrictions sanitaires imposées par la pandémie de Covid-19) et ont engagé la participation de 4 psychologues experts de la thérapie d'exposition en imagination (dont l'auteure).
Chaque patient a été en moyenne suivi pendant 80 jours (SD=88), suivant un rythme moyen d'une séance tous les 14 jours (SD=15). Nous disposons ainsi en moyenne pour chaque patient de données correspondant à 5.5 séances de thérapie (min=3, max=9), pour un total de 100 enregistrements sur les 110 séances de thérapie réalisées dans le cadre de l'étude (90.9%). Chaque enregistrement est d'une durée moyenne de 5.1 min (308 sec.).

3.4.2.1 La symptomatologie de TSPT s'amende avec les séances successives de thérapie

Avant toute analyse des données acoustiques, nous observons la symptomatologie de nos patients en début d'étude et son évolution au cours des séances successives de la thérapie.
En moyenne les patients de notre étude présentent à la première séance un score à la PCL5 supérieur au score seuil pathologique. Il est de 49.10 (SD=12.73) chez les femmes et 49.28 (SD=7.56) chez les hommes. Ce score diminue de manière

significative au fur et à mesure des séances pour atteindre un score moyen de 28 (SD=16.13) chez les femmes et 28,4 (SD=13.39) chez les hommes.

Nous testons la relation entre le score total moyen à la PCL5 et les séances successives via le modèle linéaire suivant :

$$\text{PCL\_total} \sim \text{séance} + (1|\text{patient\_id}) \tag{3.2}$$

Le résultat du modèle [séance : β =-4.86, t(80.4)= -12.02 , p= 0.0 ***] va dans le sens d'une diminution significative de la symptomatologie au fur et à mesure des séances de thérapie (-4.9 points de PCL5 par séance) (Figure 3.3).

Une observation plus fine de l'évolution de la symptomatologie, en prenant en compte les différents clusters B, C, D et E de symptômes de la PCL5 met également en évidence une diminution capturable par une tendance linéaire de la symptomatologie de chaque cluster au fur et à mesure des séances (Figure 3.3).
Pour chaque cluster et de manière séparée nous avons appliqué une analyse linéaire mixte afin de tester l'influence de la séance. Les résultats des quatre analyses suivantes sont résumés dans la table 3.3.

$$\text{PCL\_B} \sim \text{séance} + (1|\text{patient\_id}) \tag{3.3}$$

$$\text{PCL\_C} \sim \text{séance} + (1|\text{patient\_id}) \tag{3.4}$$

$$\text{PCL\_D} \sim \text{séance} + (1|\text{patient\_id}) \tag{3.5}$$

$$\text{PCL\_E} \sim \text{séance} + (1|\text{patient\_id}) \tag{3.6}$$

	Estimate (β)	DF	t-stat	p-val	Sign
PCL_B	-0.049	80.8	-8.9	0.00	***
PCL_C	-0.071	81.2	-9.6	0.00	***
PCL_D	-0.075	81.6	-8.3	0.00	***
PCL_E	-0.059	80.2	-8.7	0.00	***

TABLE 3.2 – Résumé des analyses linéaires mixtes pour chaque cluster de PCL5 en fonction des séances

Compte tenu de la régression effective de la symptomatologie, nous pouvons considérer que les comportements des marqueurs que nous mettons en évidence ci- après pourraient relever des processus en jeu dans la guérison du TSPT.

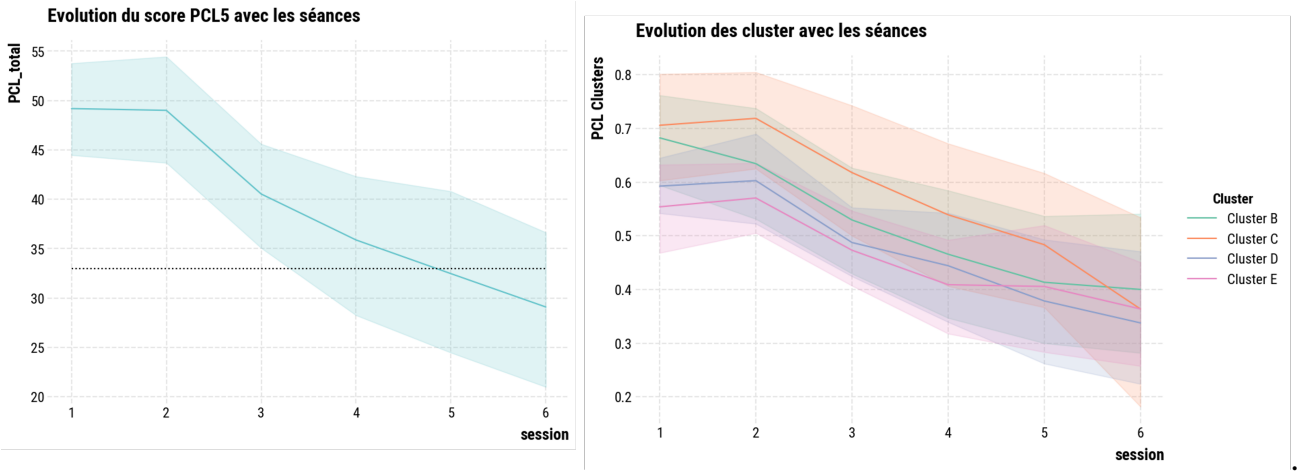


FIGURE 3.3 – Évolution de la symptomatologie de TSPT au long des séances de thérapie. A gauche, évolution du score total à l'échelle PCL5 (Weathers et al., 1993), la ligne pointillée marque le cut-off clinique de 33. A droite, évolution des symptômes de chaque cluster de la PCL5, les scores des clusters B,C,D et E ont été normalisés entre 0 (score minimum à tous les items) et 1 (score maximum).

3.4.2.2 Le pitch permettrait de discriminer entre séance pathologique et non pathologique

Afin d'examiner l'existence de biomarqueurs acoustiques du TSPT dans la voix de nos patients, nous procédons à une première analyse en séparant les séances de thérapie en séances pathologiques (pour lesquelles le score PCL5 est strictement supérieur à 32) et non pathologiques (score PCL5 ≤ 32). Nous appliquons ensuite pour chaque paramètre acoustique extrait un modèle linéaire mixte en prenant la variable « pathologique » comme régresseur et en posant le patient comme facteur aléatoire, comme le montre l'équation suivante sur la variable pitch :

$$\text{mean\_pitch} \sim \text{pathologique} + (1|\text{patient\_id}) \tag{3.7}$$

Parmi nos quatre paramètres acoustiques, seul le pitch moyen paraît être influencé de manière significative par le statut pathologique ou non des séances [`pathologique` :

β =10.93, t(8.1)= 4.27 , p= 0.003], le pitch moyen étant environ 10Hz plus élevé lors des séances pathologiques que lors des séances pour lesquelles la symptomatologie est passée sous le score pathologique.

3.4.2.3 Le pitch diminue au fur et à mesure de l'avancée de la thérapie

Puis, nous réalisons des analyses en mesures répétées sur les différentes séances consécutives de thérapies afin de mettre en évidence les facteurs susceptibles d'expliquer l'évolution des caractéristiques acoustiques extraites dans nos extraits. Pour cela, nous utilisons des modèles linéaires à effets mixtes en posant le patient comme facteur aléatoire et les caractéristiques acoustiques comme variables à expliquer. Les tests prennent en compte le sexe des patients.
Nous testons la relation entre chaque paramètre acoustique et du score moyen de PCL5 total via le modèle linéaire suivant :

$$\text{mean\_pitch} \sim \text{PCL\_total} + (1|\text{patient\_id}) \tag{3.8}$$

Seul le pitch semble présenter une évolution linéaire en fonction du score total à la PCL5. Le résultat du modèle de régression en fonction du score moyen de PCL5 total [β =0.25, t(71.8)= 2.56 , p= 0.012] va dans le sens d'une diminution significative du pitch avec la baisse symptomatologie globale. Cette baisse de pitch est de 0.25Hz par point de PCL, soit environ 5.5Hz pour une baisse moyenne de 22 points au cours de la thérapie), ce qui correspond à une baisse de 77 cents chez les hommes et de 50 cents chez les femmes (Figure 3.4).

3.4.2.4 La voix reflète les différentes trajectoires de guérison

Afin d'examiner plus en détail le processus de guérison, nous nous intéressons aux sous-catégories de symptômes inclues dans la PCL5. Comme chaque cluster n'est pas représenté par le même nombre d'items, nous normalisons les scores obtenus pour chaque cluster en fonction du nombre d'items qui le compose.
Puis, nous testons pour chaque paramètre acoustique (pitch, jitter, shimmer et NHR), l'influence des différents clusters de symptômes (B, C, D et E) en appliquant le modèle 3.9 à chacun des paramètres acoustiques :

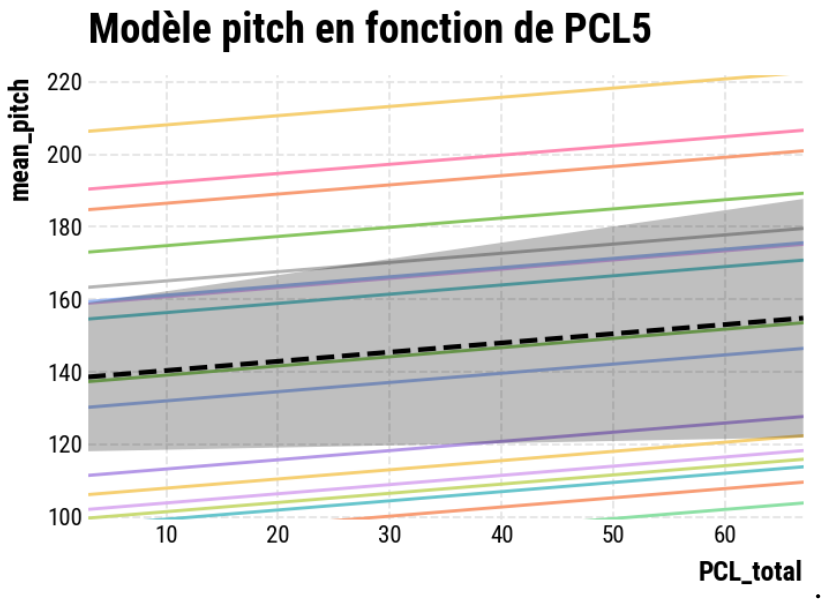


FIGURE 3.4 – Illustration des paramètres (pente et ordonnée à l'origine) du modèle (1.7) affichés sous la forme de droites parallèles. Les données ne figurent pas dans ce graphique.

$$\text{mean\_pitch} \sim \text{PCL\_B} + \text{PCL\_C} + \text{PCL\_D} + \text{PCL\_E} + \text{session} + (1|\text{patient\_id}) \quad (3.9)$$

Les analyses sur les quatre paramètres mettent en évidence une influence différente des cluster sur les paramètres acoustiques que nous résumons dans la table 3.3. La baisse des symptômes de reviviscence (cluster B) apparaît rendre compte à la fois d'une diminution de pitch [β =27.43, t(68.1)=2.63, p= .010 *] et de jitter [β =0.79, t(81.8)= 2.24 , p= 0.028 *]. La diminution des cognitions et affects négatifs (cluster D) s'observe quant à elle par une tendance d'augmentation du pitch [β =-21.3, t(67.9)= -1.85 , p= 0.068 .]. Enfin, les symptômes relatifs à l'hyperactivation neurovégétative (cluster E), en disparaissant, sont associés à une augmentation de jitter [β =- 0.96, t(81.2)= -2.33 , p= 0.022 *] , du shimmer [β =-2.77, t(77.5)= -2.17 , p= 0.033 *] et du rapport bruit sur harmonique (NHR) [β =-0.069 , t(74.67)= -1.72 , p= 0.089 .] dans la voix des patients. Enfin, aucun de nos marqueurs acoustiques ne corrèle avec le cluster C relatif aux symptômes d'évitements.

	B : Reviviscences	D : Cognitions et Affects négatifs	E : Hyperarousal
Pitch	β = 27.4, p = .010	β = 21.3, p = .068	
Jitter	β = 0.79, p = .028 *		β = -0.96, p = .017 *
Shimmer			β = -2.77 ; p = .033 *
NHR			β = -0.07 ; p = .089

TABLE 3.3 – Résumé des modèles linéaires mixtes testant l'influence des cluster de PCL5 sur les caractéristiques acoustiques. En couleurs les régressions significatives. En bleu : les relations positives, en orange : les relations négatives

3.4.3 Discussion

Nous décrivons ici une étude longitudinale menée sur 20 patients au cours des séances successives de thérapie d'exposition en imagination qui adresse la question de l'incidence du TSPT sur la voix de patients pendant qu'ils font le récit de leur expérience traumatique.
Tout d'abord et indépendamment de toute mesure acoustique, nous constatons la régression effective de la symptomatologie du TSPT au cours des séances de thérapie, témoignant de l'efficience de celle-ci dans notre cohorte. Ceci nous permet de considérer que les comportements des marqueurs acoustiques que nous mettons en évidence dans cette étude pourraient relever des processus en jeu dans la guérison du TSPT.

A ce sujet, nos résultats montrent que, parmi les quatre paramètres acoustiques que nous choisissons d'observer dans cette étude, le pitch semble être influencé par le statut pathologique/non pathologique des séances. Il est en effet significativement plus bas de 10Hz lors des séances pour lesquelles la symptomatologie post-traumatique s'est amendée (score PCL total<33) par rapport aux séances où cette dernière est pleinement active.

De façon plus précise, le design en mesures répétées de notre étude nous permet de suivre le comportement des paramètres acoustiques au long de la thérapie et d'éclairer les facteurs qui pourraient expliquer ce comportement. Nous observons que l'évolution du pitch semble être expliquée d'une part par le facteur séance (en moyenne chaque nouvelle séance de thérapie étant associée à une baisse de pitch

de 2.3 Hz) et d'autre part par le niveau de symptomatologie globale (en moyenne chaque point de PCL perdu étant associé à une baisse de pitch de 0.25Hz). Néanmoins, l'analyse conjointe de l'influence de la séance et de la symptomatologie globale au sein d'un même modèle ne permet plus d'observer d'influence du score global sur la baisse de pitch.

Ceci laisse penser que progressivement au fil des séances, un phénomène soit en jeu expliquant la réduction du pitch au cours de la thérapie, qui n'est pas entièrement expliqué par le niveau global des symptômes. Or, nous savons du consensus de l'état de l'art en matière d'expression émotionnelle vocale que des valeurs de pitch élevées sont liées à une intensité émotionnelle élevée (Bänziger et al., 2001 ; Juslin and Laukka, 2003). La diminution du pitch que nous constatons pourrait alors témoigner d'une baisse de l'intensité émotionnelle éprouvée lors de la narration de l'ET au fil des séances successives. Si l'intensité des reviviscences ne représente en effet qu'une part de la symptomatologie générale du TSPT, rappelons que ce schéma d'évolution est considéré comme essentiel pour la guérison selon la théorie du traitement émotionnel qui sous-tend cette thérapie (Foa and Kozak, 1986). Aucun autre paramètre acoustique ne semble influencé par les séances ni la symptomatologie générale.

Afin d'examiner plus finement les paramètres influençant l'évolution de la voix en cours de thérapie, nous interrogeons l'effet spécifique des différents clusters B, C , D et E rendant respectivement compte des symptômes de reviviscences, d'évitement, de cognitions et affects négatifs et enfin d'hyperéveil. A ce titre, nous observons que tous ne semblent pas influencer les comportements vocaux et que leurs effets se font de manière différentielle sur chaque paramètre acoustique ici étudié (Table 3.3).

Premièrement, la baisse des symptômes du cluster B semble expliquer à la fois les déclins du pitch et du jitter. La diminution du pitch associée à celle des symptômes de reviviscence semble cohérente avec la littérature reliant intensité émotionnelle et pitch. En effet, le cluster B, de par les items qui le constituent, rend compte en plus des symptômes d'intrusions, de l'intensité des émotions ressenties lorsque le patient est ramené à l'évènement traumatique (que ce soit en pensée ou in vivo). Pour rappel, certains des items du cluster B de la PCL5 permettent d'évaluer la détresse psychique lors de l'exposition à des stimuli évoquant l'ET (p. ex., lors de la « date

anniversaire » ou lorsque le patient entend des sons similaires à ceux entendus pendant l'événement), les réactions physiologiques marquées lors de l'exposition à des stimuli évoquant l'ET et enfin les réactions dissociatives qui consistent à agir ou souffrir comme si l'événement se déroulait de nouveau (cela va des flash-backs à une totale perte de conscience de l'environnement présent). L'interprétation de la baisse de jitter avec la baisse des symptômes de reviviscence, en plus d'être compliquée par le fait qu'elle soit faite sur de la mesure continue, est également difficile à éclairer sur le plan de l'intensité émotionnelle. En effet, les données de la littérature ne permettent pas de tirer de conclusion définitive à ce propos, si ce n'est que les manifestations de colère ont peut-être tendance à montrer plus de jitter que celles de tristesse. Certains auteurs expliquent cet état de l'art par le recours important des études sur le lien entre voix et émotions à des corpus de voix d'acteurs et proposent que le jitter puisse être un indice vocal difficile à manipuler pour les acteurs (Bachorowski and Owren, 1995).

La baisse des symptômes relatifs aux cognitions et affects négatifs (cluster D) ne semble associée qu'au pitch et à son augmentation. Ceci va plutôt dans le sens des données de la littérature liant des valeurs de pitch basses aux états de tristesse (Juslin and Västfjäll, 2008) et de dépression (Robb, 1999). Néanmoins, même si les items de ce cluster décrivent une symptomatologie partagée par le trouble dépressif, nous précisons que ce cluster n'est aucunement un outil diagnostic de dépression. Aussi, nous ne pouvons pas conclure que les niveaux de pitch bas associés à ces symptômes relèvent d'une dépression majeure sous-jacente, d'autant que celle-ci n'est pas reportée de manière spécifique à chaque séance (ceci aurait nécessité un outil diagnostique propre à la dépression). La baisse du score au cluster E semble quant à elle être associée à une augmentation des niveaux de jitter, shimmer et nhr, mais pas de pitch. Afin de donner du sens à cette relation, rappelons que ce cluster évalue les symptômes d'hyperéveil, ou encore appelé hyperactivation neurovégétative qui renvoie à l'activité du système nerveux neurovégétatif (SNV). Ce dernier est composé de deux branches sympathique et parasympathique dont l'équilibre dynamique, crucial pour le fonctionnement général de l'organisme et dans les processus de régulation émotionnelle, s'avère perturbé dans le cas du TSPT (Gillie and Thayer, 2014 ; Levenson, 2014). Sachant cela, il peut sembler étonnant d'observer que la normalisation de ce système s'accompagne d'une augmentation de trois mesures témoins de la perturbation du signal acoustique (jitter et shimmer) et du bruit (nhr) (Kreiman and Gerratt, 2005).

Or, les travaux d'Orlikoff (1989) sur la mesure du jitter précisent que la variation dans la régularité du signal est un pattern physiologique normal. Ainsi il existe en phonologie des valeurs normales attendues pour ces trois mesures, calculées à partir d'extraits de vocalisation de voyelles tenues, qui servent notamment à caractériser une voix pathologique alors généralement perçue comme soufflée, voilée ou rauque (Farrús et al., 2007 ; Kreiman and Gerratt, 2005 ; Michaelis et al., 1998). Dans notre étude, la hausse de ces mesures (calculées non pas sur des voyelles mais sur la totalité du discours) accompagnant l'affaiblissement des symptômes d'hyperactivation, pourrait témoigner d'une normalisation de ces niveaux de perturbations. Cette hypothèse va dans le sens de l'intuition clinique partagée par les soignants d'une forme de « non naturalité » entendue dans le discours des patients en début de thérapie, comme une rigidité, une absence de modulation perçue dans la voix. Toujours Orlikoff (1989) en parlant du jitter, indique que *« nous pourrions, en tant qu'auditeurs, attendre ce genre d'imperfection : plusieurs expériences menées avec des voix synthétiques ont montré qu'une variabilité trop faible d'un cycle à l'autre se traduit généralement par la perception d'une phonation non naturelle et distinctement non humaine. »*. Cependant, les données de la littérature ne renseignant pas de niveaux normaux attendus pour ces mesures dans le cas où elles sont calculés sur de la parole continue non-contrôlée (comme c'est le cas ici), nous ne pouvons statuer au sujet de cette hypothèse sans conduire d'expérience d'écoute par des tiers, de l'ordre de celle que nous ferons au chapitre suivant. En résumé, l'observation de l'influence des clusters sur les paramètres acoustiques, semble faire émerger un double mouvement opposé de ces symptômes sur la voix pendant la réactivation du souvenir traumatique en séance. D'une part l'activation émotionnelle semble être associée à des valeurs de pitch et de jitter élevées et de l'autre l'hyperactivation neurovégétative sous-tenderait des niveaux faibles de jitter, shimmer et nhr. Il nous semble alors intéressant de relier cette dichotomie apparente entre d'une part une intensité émotionnelle subjective et une hyperactivité physiologique à la proposition de l'existence d'une dissociation dans les émotions entre l'évaluation subjective d'un côté et la réactivité physiologique associée au SNV défendue par certains chercheurs dont Joseph Ledoux que nous avons évoqué précédemment pour sa contribution majeure au sujet de la peur et des troubles anxieux (LeDoux and Hofmann, 2018 ; Pine and LeDoux, 2017).

Limites :

Nous soulignons plusieurs limites relevant de la méthodologie utilisée et de l'état actuel de ce travail.

Si parmi les tendances décrites par (Schuller et al., 2007) dans le domaine "Speech and Emotion", notre étude ne se place définitivement pas dans le cas de *« la tendance à l'exploitation approfondie de l'espace des caractéristiques, qui se traduit par des centaines, voire des milliers de caractéristiques utilisées pour la classification »*, suivant notre argument initial de permettre de donner un sens clinique à nos résultats, ils s'inscrivent par contre complètement dans celle de *« la tendance vers l'utilisation de données plus naturelles et plus réelles »*. Pour autant, il nous semble que des analyses complémentaires (comme l'extraction de paramètres temporels comme le rythme de parole ou le pourcentage de pause dans le discours ou spectraux tels que les Mel-Frequency Cepstral Coefficients) permettraient de répondre de manière plus complète à la question des corrélats acoustiques associés à l'exposition en imagination à l'ET et pourraient également être réalisées aidées des techniques de l'apprentissage machine. Se poserait alors la question du nombre optimal de caractéristiques à extraire afin de ne pas perdre de vue la portée clinique de ce travail.

Par ailleurs, nous avons proposé diverses hypothèses explicatives pour nos résultats que nous pouvons nuancer. Nous avons notamment proposé d'expliquer l'augmentation du jitter associée à la baisse des symptômes d'hyperactivation neurovégétative en nous appuyant sur la littérature soulignant le caractère physiologique au sens de « normal » d'un certain niveau de variabilité. Une explication alternative à cela, pourrait se baser sur la relation connue entre pitch et jitter. En effet, nous savons notamment par le travail d'Orlikoff and Baken (1989) que le jitter est en partie expliqué par les valeurs de pitch de sorte qu'une F0 plus élevée tend à être associée à des perturbations moindres. Il faut préciser qu'il s'agit là *« d'une dépendance non linéaire observée entre la variabilité de la fréquence à court terme et la fréquence de phonation dans la voix normale. »* (Orlikoff, 1989). Dans notre corpus, la diminution du pitch au fur et à mesure des séances pourrait alors en partie être responsable de l'augmentation du jitter, les valeurs de pitch trop hautes en début de pathologie ne permettant pas des niveaux de jitter normaux. Néanmoins, de la même manière qu'évoqué précédemment, ces hypothèses restent difficiles à corroborer en l'absence de données concernant ces niveaux de normalité dans le cadre du calcul de ces paramètres sur du discours. Aussi, les hypothèses que nous avons développées afin d'expliciter l'influence différentielle

des clusters de symptômes sur la voix impliquent que nous utilisions la PCL5 d'une manière détournée de sa fonction diagnostique initiale, dans le sens où nous l'utilisons en quelque sorte pour attester de l'expérience vécue pendant l'exposition. En cela, nous en faisons un usage proche de ce qui a pu être réalisé avec la Clinician Administered Dissociative States Scale (CADSS), un entretien diagnostique standardisé qui a été utilisé dans le cadre d'une étude en imagerie cérébrale pour évaluer la réponse à un protocole d'imagerie mentale chez des patients souffrant de TSPT (Lanius et al., 2006). Les hypothèses que nous proposons alors doivent tenir compte de ce point. Néanmoins, pour certains cluster, cet usage nous semble évident, c'est le cas du cluster B explorant la dimension de reviviscence indubitablement liée à la situation d'exposition de par les items qui composent cette dimension du questionnaire.

Quoiqu'il en soit, afin de creuser les pistes d'interprétation ouvertes par nos résultats, il apparaît nécessaire d'utiliser en plus d'un outil diagnostic, une évaluation spécifique de la réponse à l'exposition en imagination. De la même manière, la possible explication de l'augmentation du jitter par un retour de l'équilibre dans la balance sympathique-parasympathique nécessite une mesure plus directe de cette activité en complément de l'évaluation qui en est faite par le cluster D de la PCL5. A cette fin, nous avons fait l'acquisition d'un corpus supplémentaire de données lors d'une extension à l'étude Traumacoustik. Cette extension de l'étude qui s'est tenue pendant une année supplémentaire nous a permis d'ajouter trois mesures en plus de celles comprises dans le protocole initial (enregistrements vocaux et mesure de PCL5) à savoir :

— la mesure de l'activité cardiaque au moyen d'un dispositif non invasif , la montre Empatica E4 (Empatica Inc, MA : Cambridge) qui permet, au moyen d'un capteur de photopléthysmographie (PPG), l'acquisition en temps réel du signal cardiaque,
— l'évaluation spécifique de la réponse à l'exposition en imagination par la Responses to Script-Driven Imagery Scale (RSDI) (Hopper et al., 2007) , une échelle en langue anglaise dont l'usage est pour l'instant limité au cadre de la recherche
— et enfin une mesure précise des émotions ressenties pendant l'exposition, via l'échelle Geneva Emotion Wheel (GEW-French) (Sacharin et al., 2012), une échelle auto-rapportée permettant au patient d'indiquer l'émotion dont il/elle

fait l'expérience en pointant une intensité pour une ou plusieurs émotions arrangée(s) sur un cercle contenant 20 familles d'émotions distinctes.

Au moment de la rédaction de cette thèse nous n'avons pas encore finalisé l'analyse de ce nouveau corpus qui nous permettra nous l'espérons en plus de trancher au sujet des différentes hypothèses alternatives abordées ici, de préciser ce que la voix peut nous dire au sujet des mécanismes en jeux pendant la thérapie d'exposition en imagination. Seule la mesure de l'activité cardiaque a commencé à être analysée. Nous présentons ces résultats préliminaires dans la prochaine section 3.5 de ce chapitre, l'ensemble des données (physiologiques et qualitatives) de cette nouvelle cohorte de patients sera traité ultérieurement et développé au sein d'un article qui sera soumis à publication.

A l'aune de cette étude, la pertinence de mesurer la voix comme indicateur de l'expérience subjective vécue pendant que le patient raconte à voix haute l'évènement traumatique vécu semble attestée.

En effet, d'après nous, l'enjeu de ce travail n'est pas d'utiliser la voix comme marqueur diagnostique, il m'apparaît que les outils cliniques dont les cliniciens disposent déjà (entretiens standardisés et autoquestionnaires) sont bien plus simples et efficaces pour remplir cet objectif diagnostique.

Par contre s'agissant de mieux comprendre ce qui se joue subjectivement en cours de thérapie, et ainsi pouvoir suivre le niveau d'intensité émotionnelle ressentie en cours d'exposition, il existe à notre avis un intérêt clinique notable à utiliser un indicateur objectif qui évolue au cours de la thérapie en parallèle du ressenti subjectif des patients, ce dernier étant souvent décrit comme difficile d'accès par les soignants pour différentes raisons telles que de faibles compétences d'*insight* ou en cas de dissociation (Hopper et al., 2007 ; Sack et al., 2012). De plus, nous savons que certains échecs de cette thérapie peuvent être liés au fait que le patient met en place des comportements d'évitement pendant l'exposition, mécanismes dont il peut avoir plus ou moins conscience (Brown et al., 2019). La possibilité d'attester de la reviviscence effective pendant l'exposition, par l'observation de marqueurs objectifs extraits dans la voix s'avèrerait alors une aide précieuse. Le cadre d'utilisation de ces marqueurs acoustiques que nous proposons dans cette thèse, à l'opposé d'une démarche de classification diagnostique qui obéit à une logique d'homogénéisation et passe sous silence les spécificités individuelles, en saisissant l'hétérogénéité des

réponses à l'exposition en imagination contribuerait à l'effort d'individualisation de la prise en charge du TSPT.

Déclaration de contribution

Nadia Guerouaou : Conception de l'étude, Collecte des données, Analyse des données : statistiques, Rédaction - préparation du projet original.
Severine Vanhoove, Stéphane Duhem, Alice Damarey, psychologues au CRP HdF : Collecte des données.
Frédérique Warembourg, cheffe de service du CRP HdF : Rédaction - révision du projet. JJ Aucouturier : Conception de l'étude, Extraction et Analyse des données acoustiques, Analyse des données statistiques, Rédaction - préparation du projet original.
Guillaume Vaiva : Conception de l'étude, Rédaction - révision du projet.

3.5 Extension TraumacoustiK

Etude longitudinale du lien entre mesures acoustiques et activité cardiaque chez les patients souffrant de TSPT pendant la thérapie d'exposition en imagination à l'évènement traumatique

Afin de mieux comprendre les mécanismes physiologiques qui sous-tendent l'évolution du pitch avec la symptomatologie du TSPT, nous avons donc réalisé une seconde étude longitudinale. Le même protocole que celui mis en oeuvre pour la première cohorte de patients y est ainsi utilisé auquel vient s'ajouter, entre autres, une mesure du volume sanguin (Blood Volume Pulse - BVP) afin d'en extraire le rythme cardiaque des patients.

3.5.1 Matériel et méthodes

A l'exception de ces mesures supplémentaires, le protocole est strictement le même. Par conséquent, nous décrivons uniquement les éléments de méthodologie ne relevant pas de la même procédure que pour la précédente recherche.

3.5.1.1 Encadrement réglementaire de l'étude : aspects éthique et légaux

Cette extension de l'étude Traumacoustik a fait l'objet d'une demande d'avis de modification substantielle (MS) auprès du CPP HdF1 qui, après avoir revu l'intégralité des modifications au protocole ainsi que l'ensemble des mises à jour concernant la totalité des documents qui l'accompagnent (CRF papiers, Lettre d'information et formulaire de non opposition) a donné son accord le 8 janvier 2021 (Nos références CPP Ile de France 1 - NUMERO DOSSIER : 2020 déc. MS294 – avis final). Les MS, malgré l'ajout d'une mesure de l'activité cardiaque n'ayant pas changé le statut de l'étude, celle-ci reste considérée comme une étude RIPH de Catégorie 3.

3.5.1.2 Participants

La population de notre étude est constituée de 14 patients (10 femmes) reçus au Centre Régional des Psychotraumas Hauts-de-France (CHRU de Lille) (dirigé par Dr Frédérique Warembourg) et recrutés selon les mêmes critères que la précédente étude.

3.5.1.3 Procédure

La procédure est la même que pour la précédente étude à l'exception de l'ajout de la mesure de l'activité cardiaque à partir de la première visite V1 jusqu'à la fin de la dernière séance d'exposition en imagination.

3.5.1.4 Mesure indirecte de l'activité cardiaque :

En plus de l'enregistrement des productions vocales des patients réalisés au cours des séances d'exposition en imagination, nous enregistrons leur activité psychophysiologique cardiaque au moyen du bracelet Empatica E4. Il s'agit d'un dispositif médical certifié CE portable qui permet l'acquisition en temps réel des signaux physiologiques. Nous faisons ici le choix d'utiliser le dispositif le moins invasif possible afin de ne pas interférer avec la thérapie d'exposition. Les performances du bracelet E4 ont été validées dans plusieurs études (McCarthy et al., 2016) et concernent divers domaines de recherche tels que la surveillance du sommeil (Onton et al., 2016), la sécurité au volant (Kundinger et al., 2020) et l'évaluation de la régulation émotionnelle (Matsubara et al., 2016). Dans notre protocole, à chaque séance, le bracelet est placé au

poignet du patient juste avant de commencer l'enregistrement vocal. Au moment où le patient commence à parler, il appuie sur un bouton présent sur la montre pour marquer le début de la phase d'exposition et appuie sur le même bouton lorsqu'il a fini.

3.5.1.5 Analyses acoustiques

Les analyses acoustiques sont réalisées suivant la même méthodologie que dans le précédent chapitre. Ce manuscrit ne rendons compte que des données concernant le pitch.

3.5.1.6 Analyses du signal BVP : extraction du RC

Dans cette étude, le RC nous est donnée via une mesure indirecte appelée *Blood Volume Pulse* (BVP) mesurée au moyen d'un PPG implémenté au sein du bracelet `Empatica E4`.
Le BVP est un signal quasi-périodique constitué de pics d'impulsions successifs, générés par l'activité de pompage du cœur. La méthode de photopléthysmographie (PPG) utilisée pour enregistrer ce BVP repose quant à elle sur l'utilisation d'une source lumineuse et d'un récepteur pour mesurer la quantité de lumière absorbée ou réfléchie par la peau humaine. En effet, le volume sanguin dans les tissus cutanés affecte la manière dont la peau réagit à la lumière, ce qui permet de déduire la fréquence cardiaque. La PPG présente l'avantage d'être une méthode simple et moins invasive que l'électrocardiogramme (ECG), qui mesure l'activité électrique du cœur pour déterminer la fréquence cardiaque. Le bracelet E4 calcule les intervalles entre les battements (*inter-beat interval*, IBI) à partir du signal BVP que nous n'avons pas choisi d'utiliser pour nos analyses. De fait, l'IBI fourni par le bracelet E4 est fortement filtré à l'aide d'un algorithme propriétaire. Les données PPG sont analysées en Python (vers. 3.8.10) en utilisant la procédure par défaut du module `heartpy` (vers. 1.2.7, © van Gent, 2019-2023) ci après :

— Tout d'abord, nous appliquons un filtre passe-bande (entre 0.7Hz et 3.5Hz, d'ordre 3).

— A chaque point temporel, nous définissons un seuil pour détecter les pics élevés en calculant la moyenne d'une fenêtre de 750 ms centrée sur ce point (technique du « moving average »).
— Toutes les données supérieures à ce seuil sont définies comme des pics ; pour chaque pic, nous considérons la position du maximum comme la position de la systole (ou "pic R").
— Les intervalles RR (intervalle de temps entre deux pics R) sont calculés entre les pics R successifs, puis convertis en minutes, moyennés et inversés pour obtenir une valeur moyenne de BPM pour la séquence analysée.
— Si le BPM obtenu est inférieur à 40 ou supérieur à 180, ou si l'écart-type du RR est supérieur à 0.1 ms, cette procédure est répétée avec un seuil progressivement plus élevé, augmenté par paliers de 5 % jusqu'à ce que ces conditions soient remplies.
— Enfin, les pics R sont supprimés si leurs intervalles RR précédents sont des valeurs aberrantes définies comme +/- 30 % de l'intervalle RR moyen. La valeur finale du BPM est calculée comme ci-dessus, en utilisant cette liste finale de pics.

Ces analyses nous permettent d'obtenir une valeur de bpm moyen pour chaque séance de thérapie.

3.5.1.7 Analyses statistiques

Les analyses statistiques que nous présentons sont faites en Python, avec les modules `seaborn` (© M. Waskom, 2021-2023) `pingouin` (© R. Vallat, 2018-2023) et `pymer` (© E. Jolly, 2017-2023). Pour les analyses de régression, `pymer` utilise le module `lmer` qui estime le modèle testé à l'aide de la méthode de vraisemblance restreinte ou résiduelle (acronyme anglais REML) et les p values ainsi que les degrés de libertés (DF) sont estimés par l'approximation de Satterthwaite.
Les données ont été analysées en utilisant des modèles de régression linéaire mixtes (*linear mixed models*, LMM) pour lesquels nous présentons l'équation du modèle (dans sa syntaxe *R* (Banse and Scherer, 1996)). La pertinence de chaque paramètre du modèle sera illustrée par les résultats du test T de Student de nullité statistique du coefficient associé à chaque paramètre (H0 : $\beta = 0$, contre H1 : $\beta \neq 0$) sous la forme de la valeur du coefficient β d'estimation du paramètre, la valeur du t et sa p-value.

Les analyses suivantes sont réalisées suivant un plan expérimental intra-patient.

3.5.2 Résultats

Les données que nous présentons ont été collectées entre Février 2021 et Août 2022 et ont engagé la participation de 4 psychologues experts de la thérapie d'exposition en imagination (dont moi-même).
Chaque patient a été en moyenne suivi pendant 80 jours (SD=88), suivant un rythme moyen d'une séance tous les 14 jours (SD=15). Nous disposons ainsi en moyenne pour chaque patient de données correspondant à 5.9 séances de thérapie (min=1, max=9), pour un total de 80 enregistrements sur les 88 séances de thérapie réalisées dans le cadre de l'étude (90.9%). Chaque enregistrement est d'une durée moyenne de 9.10 min. (595 sec.)

Afin de pouvoir examiner les liens entre le pitch et l'activité cardiaque au cours de la thérapie d'exposition en imagination, il nous semble important de vérifier deux points essentiels. Premièrement que la thérapie fonctionne bien, c'est à dire que les symptomes régressent avec les séances puis que nous observerons le comportement du pitch et sa capacité à discriminer les séances pour lesquelles la symptomatologie post traumatique est active de celles où elle s'est amendée.

3.5.2.1 La symptomatologie de TSPT s'amende avec les séances successives de thérapie

Nous commençons par observer la symptomatologie de nos patients en début d'étude et son évolution au cours des séances successives de la thérapie.
En moyenne les patients de notre étude présentent à la première séance un score à la PCL5 supérieur au score seuil pathologique. Il est de 41.80 (SD=11.10) chez les femmes et 44.75 (SD=6.95) chez les hommes. Ce score diminue de manière significative au fur et à mesure des séances pour atteindre un score moyen de 20.30 (SD=17.29) chez les femmes et 32.75 (SD=8.09) chez les hommes.
Nous testons la relation entre le score total moyen à la PCL5 et les séances successives via le modèle linéaire suivant :

$$\text{PCL\_total} \sim \text{séance} + (1|\text{patient\_id}) \tag{3.10}$$

Le résultat du modèle [`séance` : β =-3.54, t(72.7)= -8.16 , p= 0.0] va dans le sens d'une diminution significative de la symptomatologie au fur et à mesure des séances de thérapie (-3.5 points de PCL5 par séance).

3.5.2.2 Le pitch permettrait de discriminer entre séance pathologique et non pathologique

Comme précédemment nous analysons le comportement du pitch en séparant les séances de thérapie en séances pathologiques (pour lesquelles le score PCL5 est strictement supérieur à 32) et non pathologiques (score PCL5 ≤ 32). Nous appliquons ensuite le modèle linéaire mixte en prenant la variable « pathologique » comme régresseur et en posant le patient comme facteur aléatoire suivant :

$$\text{mean\_pitch} \sim \text{pathologique} + (1|\text{patient\_id}) \tag{3.11}$$

Les résultats montrent une tendance en faveur de l'influence du statut pathologique ou non des séances [`pathologique` : β =5.47, t= 2.035, p= 0.072], sur le pitch, celui étant environ 5Hz plus élevé lors des séances pathologiques que lors des séances pour lesquelles la symptomatologie est passée sous le score pathologique.

3.5.2.3 Le pitch diminue au fur et à mesure de l'avancée de la thérapie

Nous testons d'une part la relation entre le pitch et le score moyen de PCL5 total et le pitch et les séances de l'autre via les modèles linéaires mixtes suivant :

$$\text{mean\_pitch} \sim \text{PCL\_total} + (1|\text{patient\_id}) \tag{3.12}$$

$$\text{mean\_pitch} \sim \text{session} + (1|\text{patient\_id}) \tag{3.13}$$

Le pitch semble présenter une évolution semblable à celle que nous avions observée dans la précédente étude. De fait, le résultat du modèle de régression en fonction du score moyen de PCL5 total [β = 0.25, t= 2.63 , p= 0.011] va dans le sens d'une

diminution significative du pitch avec la baisse symptomatologie globale. Cette baisse de pitch est du même ordre que celle que nous avions constatée au sein de la cohorte précédente (β = 0.25). Le pitch semble également diminuer au fur et à mesure des séances de manière significative [β = -1.05, t= -2.02 , p= 0.047]
Les résultats de notre nouvelle cohorte concernant l'efficacité de la thérapie et le comportement du pitch avec la guérison semblent répliquer ceux que nous avions observés au sein de notre première cohorte de patients. Cela posé, nous nous intéressons au comportement du RC en lien avec la guérison, puis enfin nous adresserons l'hypothèse qu'il puisse contribuer à expliquer celui du pitch.

3.5.2.4 Le RC ne semble pas évoluer de manière significative avec la progression de la thérapie

Afin d'examiner le comportement du RC au cours de la thérapie, nous testons d'une part la relation entre le bpm et le score moyen de PCL5 total et le bpm et les séances de l'autre via les modèles linéaires mixtes suivant :

$$\text{mean\_bpm} \sim \text{PCL\_total} + (1|\text{patient\_id}) \tag{3.14}$$

$$\text{mean\_bpm} \sim \text{session} + (1|\text{patient\_id}) \tag{3.15}$$

Les résultats des deux modèles ne permettent d'associer l'évolution du rythme cardiaque ni avec le niveau global de symptômes pour le premier modèle [β = 0.041, t= 0.19, p= 0.85] ni avec les séances de thérapie pour le second [β = -0.38, t= -0.32, p= 0.75].

L'évolution du pitch serait sous la double influence de la symptomatologie totale et du RC

3.5.2.5 Mise en évidence d'une relation pitch-RC

Enfin, nous testons l'effet conjoint de la symptomatologie post traumatique globale et de l'activité cardiaque sur l'évolution du pitch au sein d'un unique modèle linéaire mixte :

$$\text{mean\_pitch} \sim \text{PCL\_total} + \text{mean\_bpm} + (1|\text{patient\_id}) \tag{3.16}$$

Les résultats de ce modèle indiquent que le comportement du pitch pourrait être à la fois expliqué par la symptomatologie globale [β = 0.25, t= 2.61, p= 0.012], comme nous l'avons vu la baisse de pitch s'accompagne de celle des symptômes et également par le RC du patient dans le sens où l'augmentation du RC induirait une baisse de pitch [β = -0.29, t= -2.84, p= 0.007 **] (Figure 3.5).

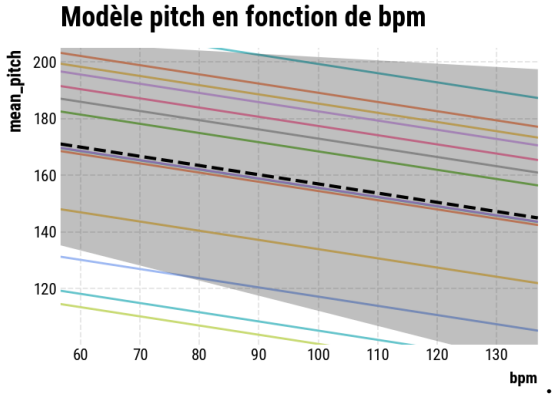


FIGURE 3.5 – Illustration des paramètres (pente et ordonnée à l'origine) du modèle (1.15) affichés sous la forme de droites parallèles. Chaque couleur représente un patient, ici les femmes sont majoritairement représentées (avec pitch > 150).

3.5.3 Discussion

Les premières analyses menées sur cette deuxième cohorte longitudinale de patients semblent attester de la pertinence de l'étude du pitch dans le cadre de la thérapie d'exposition en imagination. Les résultats sur ce second corpus confirment l'influence du statut pathologique ou non des séances sur le pitch et la baisse de celui-ci avec la régression de la pathologie. Concernant notre nouvelle mesure de l'activité cardiaque, nos analyses du RC ne permettent pas de montrer d'influence de la symptomatologie globale ni des séances successives de thérapie. Des analyses complémentaires de ce signal gagneraient à être explorées. En particulier, l'extraction de la variabilité de la fréquence cardiaque (VFC) (Heart Rate Variability HRV en anglais) à partir du signal PPG, une mesure qui témoigne des variations de temps normales entre les battements cardiaques, nous permettrait de mieux saisir l'implication différentielle des différentes

branches du SNV dont l'équilibre pourrait évoluer au cours de la thérapie sans que le seul RC puisse en rendre compte. Ainsi, *« alors que la fréquence cardiaque peut être stable, le temps entre deux battements cardiaques peut être très différent et sa valeur informative est plus importante »*pour rendre compte de l'équilibre sympatho-vagal (Marsac, 2013).

Enfin, la baisse du pitch que nous constatons de nouveau au sein de cette cohorte avec la rémission de la symptomatologie semble être également associée au rythme cardiaque du patient. L'association entre le pitch et le RC que l'analyse de cette nouvelle cohorte met en exergue semble s'inscrire dans un ensemble de données corrélationnelles liant ces deux variables dans la littérature chez l'individu non malade (Schuller et al., 2014). Cependant, le fait que l'augmentation du pitch soit associée à une baisse du RC n'est pas retrouvée dans la littérature (Usman et al., 2021). Une hypothèse serait alors que la perturbation de la balance sympathique parasympathique décrite chez ces patients (Thayer et al., 2012) puisse venir perturber cette relation pitch-RC. Afin donc de mieux comprendre le lien entre pitch et RC nous nous proposons de l'explorer chez des individus non malades qui constitue un cadre permettant de contrôler pour le mieux les facteurs pouvant influencer ces deux mesures.

Déclaration de contribution

Nadia Guerouaou : Conception de l'étude, Collecte des données, Analyse des données : statistiques, Rédaction - préparation du projet original.
Ludivine Gautier, Severine Vanhoove, Stéphane Duhem, psychologues au CRP HdF : Collecte des données.
Frédérique Warembourg, cheffe de service du CRP HdF : Rédaction - révision du projet.
Paul Maublanc : Analyse du signal PPG, Extraction et Analyse des données acoustiques, Analyse des données statistiques.
JJ Aucouturier (co-encadrant) : Conception de l'étude, Extraction et Analyse des données acoustiques, Analyse des données statistiques, Rédaction - préparation du projet original.
Guillaume Vaiva (co-encadrant) : Conception de l'étude, Rédaction - révision du projet.

3.6 Conclusion du chapitre

Les résultats précédents et en particulier le fait qu'il existe bien des paramètres dans la voix qui suivent l'évolution du TSPT pendant l'exposition en imagination à l'ET nous confortent dans l'intuition que des transformations simples de la voix comme une augmentation ou diminution de pitch pourraient, selon les contextes, reproduire des comportements vocaux à haute signification personnelle : ici, un filtre de pitch sur la voix d'un patient pourrait « simuler » le passage d'un statut malade ou guéri ainsi que l'intensité émotionnelle ressentie par le patient lors du récit de l'évènement traumatique (en lien avec le cluster B). Rappelons que nous utilisons le mot « filtre » dans une acception particulière ici - qui renvoie à la manipulation informatique de la voix - afin de mettre en exergue et de pouvoir s'emparer plus facilement des usages potentiels de ce nouvel objet. Dans les faits, quand nous parlerons de « filtre de pitch » dans la suite de notre travail, nous renverrons à une transformation de pitch réalisée dans cette thèse au moyen d'un algorithme de vocodeur de phase.

Dès lors, nous souhaitons pour la suite de ce travail, examiner le filtre du pitch comme cas d'étude d'une forme d'anthropologie cognitive de l'objet filtre vocal et tenter d'apporter des éléments de réponse à la question que nous avons introduite en début de manuscrit : quelles seraient les conséquences sur le plan cognitif de l'utilisation des technologies de transformation de l'expression émotionnelle dans la voix pour les interactions sociales ? A cette fin, nous interrogerons les effets d'un tel filtre sur nos perceptions. Le chapitre 4, dans la suite directe des résultats de TraumacoustiK, questionnera les conséquences d'un tel filtre de pitch, appliqué à la voix du patient, sur la perception des soignants. Enfin, au chapitre 5, nous élargirons notre étude des effets de la manipulation de pitch à l'individu non malade. Nous étudierons alors le lien entre le pitch et l'activité cardiaque et l'incidence du filtre de pitch sur les inférences qui peuvent être faites au sujet du rythme cardiaque d'un locuteur à partir de la voix.

4. L'évaluation psychopathologique au prisme du filtre vocal

Volet perceptif de l'effet du filtre de pitch

4.1 Introduction

Ce chapitre ouvre la question des effets potentiels du filtre de pitch sur les inférences que les individus font naturellement au sujet de leur interlocuteur dans le cadre des interactions sociales. Plus précisément, nous étudions ici le contexte spécifique de l'interaction soignant-patient et la possibilité des soignants à percevoir dans la voix des indices de l'état de gravité du patient.

Les cliniciens observent depuis longtemps que les personnes souffrant de troubles psychiatriques présentent des modifications de la parole (Breiman, 2001 ; Newman and Mather, 1938) et il est communément admis, sans que cela ne soit objectivé à notre connaissance, qu'ils utilisent dans leur routine leurs impressions perçues concernant la qualité de la voix des patients comme éléments d'évaluation de l'état psychique du patient (Révis, 2013). Ce constat renvoie à la question plus large du raisonnement clinique permettant la pose de diagnostic, et des indices sur lesquels il repose. Le raisonnement clinique en médecine a beaucoup préoccupé les chercheurs depuis le travail inaugural réalisé par Elstein et al. (1978) à la fin des années 1970. Malgré plusieurs décennies de recherche sur ce thème, comprendre le raisonnement qui étaye la prise de décision médicale représente toujours un défi extraordinaire, car les processus cognitifs qui la sous-tendent sont, par définition, inobservables et en partie activés de manière inconsciente, ce qui explique la difficulté qu'ont les médecins à les décrire. De fait, la pose de diagnostic requiert, en plus des stratégies dites de raisonnement

analytiques (typiquement conscientes/contrôlées), des stratégies de raisonnement non analytiques (possiblement inconscientes ou automatiques) (Elstein, 1999 ; Eva, 2005). Ces processus de raisonnement non analytiques « *reposent sur les capacités du praticien à reconnaître sans effort conscient une configuration caractéristique de données contextuelles et de signes cliniques évoquant très fortement un ou plusieurs diagnostics* » (Pelaccia et al., 2011). Ces connaissances tacites seraient généralement acquises au cours de l'observation et de la pratique et comprennent les expériences antérieures (Epstein, 1999). Au sujet du processus d'acquisition de ces connaissances, Barrows and Feltovich (1987) élaborent une théorie du « script de pathologie » explicitée dans la citation suivante. « *A partir des présentations didactiques, des jeux de rôle, des analyses de cas et de l'exposition clinique, les novices intègrent des réseaux d'informations, des liens associatifs et des souvenirs de rencontres réelles avec des patients pour former des grappes d'informations uniques pour chaque diagnostic.* ». Ces ensembles complexes de connaissances sont alors appelés « script de pathologie » par les auteurs qui précisent que ces réseaux de connaissances adaptés aux tâches cliniques se développent au fil de l'expérience et fonctionnent de manière autonome et non-consciente (Charlin et al. (2000),Charlin et al. (2007)).

L'usage potentiel par les soignants d'indices vocaux nous paraît pouvoir relever de ces processus non analytiques et pourrait ainsi faire partie de ce « script de pathologie » dont le fonctionnement renvoie quelque peu il nous semble au « modèle » de la théorie du traitement prédictif. Cependant, à notre connaissance, peu d'études se sont employées à tester expérimentalement cette possibilité. Parmi elles, le travail de Boidron et al. (2016) interroge l'influence du niveau de dominance véhiculé par la voix de patients fictifs - liée à la dimension de masculinité vocale - sur la décision médicale en contexte de centre d'appel médical. En manipulant artificiellement la voix des patients fictifs, les chercheurs ont montré que les appelants dont la voix était perçue comme indiquant une domination physique (c'est-à-dire ceux dont la voix avait une fréquence fondamentale basse et un écart formantique faible) ont obtenu un niveau de réponse plus élevé, une meilleure évaluation de l'urgence médicale et une attention plus longue de la part des médecins que les appelants ayant des besoins médicaux strictement identiques mais dont la voix indiquait une domination physique plus faible. De surcroît, si l'effet était important pour les médecins participants, il était pratiquement inexistant lorsque les appels étaient traités par des opérateurs

téléphoniques n'ayant pas reçu de formation médicale. Ce résultat démontre l'influence des signaux vocaux sur la décision médicale et il nous semble plaider en faveur du fait que ces indices puissent faire partie d'un ensemble de connaissances acquises par l'expérience médicale.

Cela posé, et suite aux résultats de l'étude TraumacoustiK, nous nous intéressons à l'importance spécifique du pitch en tant qu'indice vocal potentiellement utilisé par les soignants et nous souhaitons tester la possibilité qu'un filtre de pitch puisse influencer la perception par les soignants de l'état d'avancement d'un patient en cours de thérapie d'exposition en imagination.
A cet effet, nous développons le protocole expérimental suivant.

Parmi les enregistrements de notre première étude TraumacoustiK, nous sélectionnons des paires d'extraits de voix de patients correspondant au même moment du script traumatique, enregistré lors de séances situées en début et en fin de thérapie. A partir de ces extraits, nous créons des extraits « transformés » en appliquant différents niveaux de transformation de hauteur, grâce à un algorithme de pitch shifting . Les transformations opérées sur ces extraits vocaux sont faites au regard des connaissances que l'étude TraumacoustiK nous apporte au sujet de l'évolution du pitch au long de la thérapie d'exposition en imagination, à savoir que le pitch baisse au fur et à mesure des séances de thérapie et il est significativement plus bas (en moyenne de 10Hz) chez les individus pour lesquels la symptomatologie post traumatique s'est amendée. Ainsi, nous appliquons une diminution de pitch sur des extraits sélectionnés en début de thérapie pour donner l'impression que le patient est plus avancé dans le traitement (dit autrement et de manière schématique créer un effet « guéri ») et une augmentation de pitch sur des extraits correspondant aux séances finales de prise en charge pour donner l'impression que le patient est plus en amont de thérapie (toujours de manière schématique créer un effet « malade »). Pour ces extraits « transformés », seul le pitch est modifié. Le reste des informations contenues dans la voix porté par les divers marqueurs acoustiques - tels que le jitter, le shimmer et le NHR mis en évidence dans l'étude TraumacoustiK ou par encore d'autres indices que nous n'avons pas étudié comme le rythme de la voix, les pauses, le timbre - ne font l'objet d'aucune transformation. Ces extraits naturels et transformés sont présentés deux à deux à un groupe de soignants spécialistes du TSPT. Ceux-ci ont pour consigne de les écouter et

évaluer si dans l'extrait numéro 2 le patient semble aller mieux ou moins bien que dans l'extrait numéro 1.

Cette étude adresse plusieurs questions. Tout d'abord, elle examine pour la première fois à notre connaissance de manière objective la capacité des soignants à percevoir dans la seule voix du patient l'amélioration de son état en cours de thérapie d'exposition en imagination et l'effet de leur expertise sur cette capacité. Ensuite, en manipulant le pitch dans les extraits présentés, elle questionne la part jouée spécifiquement par ce paramètre dans le jugement que les soignants pourraient faire au sujet de l'état psychique d'un malade. Enfin, sur un plan plus « applicatif », ce travail examine la possibilité qu'à le « filtre pitch » à influencer le jugement des thérapeutes dans le sens de la relation que nous avons pu mettre en évidence au sein de la cohorte TraumacoustiK : à savoir un « effet guéri » associé à une baisse de pitch et un « effet malade » associé à une augmentation de pitch.

4.2 Matériel et Méthodes

4.2.1 Création des stimuli

Parmi le corpus d'enregistrements de TraumacoustiK, nous avons sélectionné 7 paires de courts passages dans les scripts de patients, chacune étant constituée d'un enregistrement au début de la thérapie (séance 1 ou 2), que nous appelons extrait S0 et d'un enregistrement du même passage à la fin de la thérapie (séance 5 ou 6) que nous appelons extrait Sf.
S0 et Sf correspondent au même passage du script respectivement « lu » en séance de thérapie quand le patient est malade (PCL5 $>$ ou $=$ 32 en S0) et quand le patient est guéri (PCL5 $<$ 32 en Sf).
La durée des passages est courte (M=5.2s, SD=1.9), correspondant à 1 ou 2 phrases (ex. "*Je me sens toujours extrêmement fatiguée, et me rendors assez rapidement*") et les paires sont extraites des enregistrements de 3 patients (patiente 01-LM : 3 paires ; patient 04-LG : 2 paires ; patiente 07-KT : 2 paires).
Conformément aux tendances analysées dans le corpus TraumacoustiK, le pitch moyen en début de thérapie pour nos paires était plus élevé qu'en fin de thérapie, d'environ M=14.2 Hz (SD=7.63). La figure 4.1 montre un exemple de paires d'enregistrements

(patiente 01-LM, extrait 1), avec un pitch moyen en séance 2 de 185.09 Hz, et un pitch moyen en séance 5 de 167.73Hz. Le pitch des extraits est donné dans la Table 4.2.

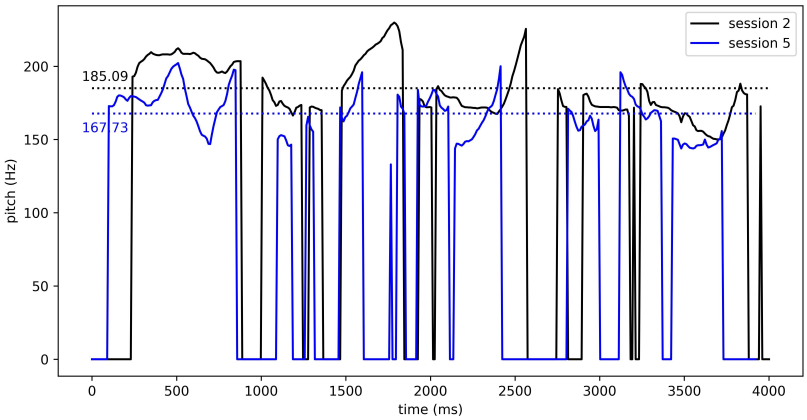


FIGURE 4.1 – Exemple de paire d'enregistrements pour le patient 01-LM extrait 1, avec un pitch moyen en séance 2 de 185.09 Hz, et un pitch moyen en séance 5 de 167.73Hz.

Pour chaque enregistrement, afin de tester l'effet du pitch, nous avons ensuite produit plusieurs versions transformées avec la procédure suivante :

— au sein de chaque paire, nous transformons l'enregistrement de début (S0) de façon à réduire son pitch en direction de celui de l'enregistrement de fin (Sf) (ex. l'enregistrement 01LM2_extr1 est transformé pour aller du pitch 185Hz à 167Hz) et, de façon symétrique, nous transformons l'enregistrement de fin (Sf) de façon à augmenter son pitch en direction de celui de l'enregistrement de début (S0) (ex. l'enregistrement 01LM5_extr1 est transformé pour aller du pitch 167Hz à 185Hz).

— Trois niveaux de transformation sont alors appliqués en augmentant ou en diminuant le pitch : 50% (on rapproche l'enregistrement transformé de l'enregistrement cible à 50% de l'écart de pitch entre les deux), 100% (on transforme au même niveau que l'enregistrement cible) et 150% (on dépasse le pitch de l'enregistrement cible de 50%). Par exemple, l'enregistrement 01LM2_extr1 est transformé pour aller du pitch 185Hz à 176Hz (50%), 167Hz (100%, correspondant au pitch de 01LM5_extr1) et 159Hz (150%).

	subject	session	extract_nb	file_name	mean_pitch	duration
0	01-LM	2	1	01-LM2_extr1.wav	185.097838	4.008299
1	01-LM	5	1	01-LM5_extr1.wav	167.730948	3.928844
2	01-LM	2	4	01-LM2_extr4.wav	186.592127	8.307211
3	01-LM	5	4	01-LM5_extr4.wav	178.679996	7.271565
4	01-LM	2	5	01-LM2_extr5.wav	181.588766	3.935011
5	01-LM	5	5	01-LM5_extr5.wav	173.365664	4.309751
6	04-LG	1	1	04-LG1_extr1.wav	116.204170	3.836916
7	04-LG	6	1	04-LG6_extr1.wav	89.442010	2.841043
8	04-LG	1	2	04-LG1_extr2.wav	125.863704	5.139683
9	04-LG	6	2	04-LG6_extr2.wav	104.668385	8.255873
10	07-KT	1	2	07-KT1_extr2.wav	200.345665	6.997506
11	07-KT	5	2	07-KT5_extr2.wav	190.131027	7.034104
12	07-KT	1	3	07-KT1_extr3.wav	194.840941	3.837324
13	07-KT	5	3	07-KT5_extr3.wav	187.122087	3.126304

FIGURE 4.2 – Table des pitch des extraits utilisés pour l'expérience

Pour chaque patient nous obtenons donc, en plus des deux extraits naturels S0 et Sf, 6 extraits transformés respectant le format suivant : S0-50%, S0-100%, S0-150% d'une part et Sf+50%, Sf+100%, Sf+150% d'autre part.
Toutes les transformations de pitch sont réalisées avec une algorithme de vocodeur de phase (implémentation de la toolbox voimooo, © Alta Voce, 2020-2023).

4.2.2 Procédure

Les extraits sont présentés par paire chacune constituée pour un essai d'une version de S0 (naturelle ou transformée) et d'une version de Sf (naturelle ou transformée). Les paires présentées sont ainsi : S0 _ Sf, S0 -50 % _ Sf , S0 -100 % _ Sf, S0 -150 %_Sf, S0 _ Sf+50%, S0 _ Sf+100%, S0 _ Sf+150% et donnent lieu à 49 essais (soit 7 combinaisons des 7 paires sélectionnées pour cette étude) . Ces 49 stimuli sont organisés en sept blocs de sept items. Les sept items sont choisis au préalable afin que chacun contienne exactement une fois chaque phrase et une fois chaque combinaison de pitch, de façon à éviter une comparaison trop évidente de plusieurs transformations de la même phrase. Les blocs sont présentés aléatoirement, et les stimuli eux-mêmes sont présentés dans un ordre aléatoire au sein de chaque bloc. 26 items (fixes) sont

présentés avec la version d'origine de la phrase (avec pitch modifié ou pas) à gauche. 23 items (fixes) sont présentés avec la version finale de la phrase (avec pitch modifié ou pas) à gauche.
Pour chaque essai, la tâche du participant est d'écouter les deux extraits, puis d'évaluer l'état du patient dans l'extrait 2 par rapport à l'extrait 1 en positionnant un curseur sur une échelle allant de -5 (le patient va beaucoup moins bien dans l'extrait 2 que dans l'extrait 1) à 100 (le patient va beaucoup mieux dans l'extrait 2 que dans l'extrait 1), en passant par 0 (pas de différence dans l'état du patient entre les extraits 1 et 2) comme illustré dans la Figure 4.3.

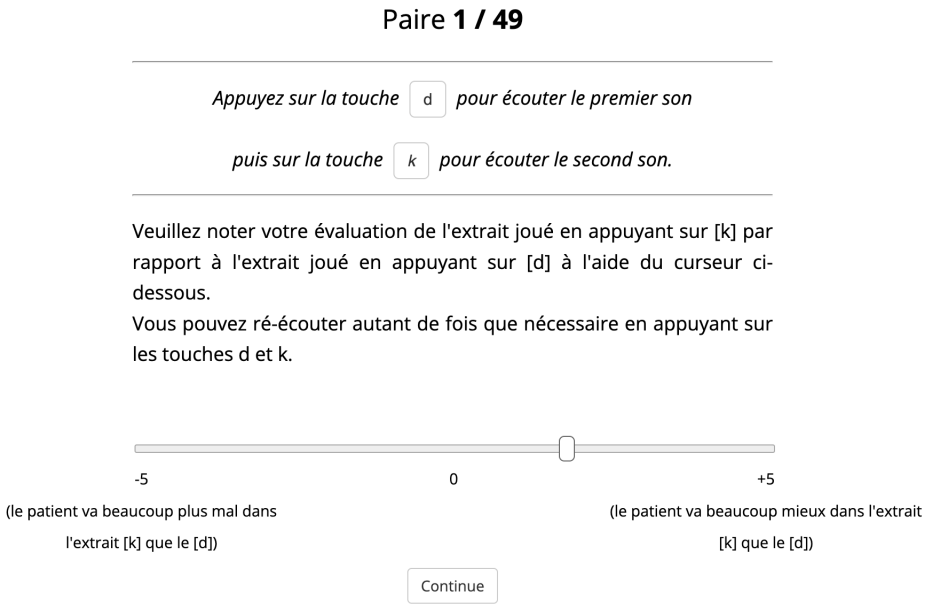


FIGURE 4.3 – Illustration de la tâche d'évaluation comparative de l'état de gravité du patient à l'écoute de deux extraits appelés d et k réalisée par les soignants

Cette étude est réalisée en ligne. L'expérience est codée en javascript, en utilisant la bibliothèque de création d'expériences comportementales en ligne jsPsych (https://www.jspsych.org/ - (de Leeuw et al., 2023)) et hébergée sur le serveur MindProbe (https://mindprobe.eu/) exécutant le logiciel libre de gestion d'expériences JATOS (Lange et al., 2015).

4.2.3 Description des variables et hypothèses associées

Notre variable dépendante correspond au positionnement de la réponse des soignants sur l'échelle d'évaluation proposée et va de 0 qui correspond au -5 (le patient va beaucoup moins bien en Sf qu'en S0) à 100 (le patient va beaucoup mieux en Sf qu'en S0), en passant par 50 qui correspond au 0 (pas de différence dans l'état du patient entre Sf et S0).
Nous souhaitons étudier l'effet de deux variables indépendantes sur cette réponse.

1. L'expertise du soignant dans la thérapie d'exposition en imagination (variable `exposition en imagination`). En effet, certains soignants utilisent cette forme de thérapie dans leur prise en charge du TSPT (`exposition en imagination=oui`) et d'autres non (`exposition en imagination=non`). Nous souhaitons observer si ce niveau d'expertise influence la performance à la tâche d'évaluation de Sf par rapport à S0.
2. La variable que nous manipulons est appelée `percent transformed` et renvoie aux 7 niveaux de transformation possibles appliqués aux extraits (0, +/- 50, +/- 100, +/- 150). Nous souhaitons étudier l'effet de l'application des filtres sur S0 et Sf sur la différence perçue entre l'état des patients dans les deux extraits présentés. Dès lors, nous faisons l'hypothèse que si les soignants fondent leur jugement sur le pitch, alors d'une part quand la baisse de pitch est appliquée sur S0, la différence perçue entre S0 manipulé et Sf devrait être moindre qu'en condition S0 et Sf naturels et d'autre part, quand la hausse de pitch est appliquée sur Sf, la différence perçue entre S0 naturel et Sf au pitch augmenté devrait également être moindre par rapport aux essais comparant deux extraits naturels. Nous faisons l'hypothèse que ces diminutions de différence perçue seront d'autant plus importante que le niveau de transformation opérée sur l'extrait sera grand.

4.2.4 Participants

Les participants de cette étude sont inclus de mai à juillet 2023. Il s'agit de 17 thérapeutes participent à l'étude (4 médecins psychiatres, 2 infirmiers-psy, 10 psychologues, une psychomotricienne), provenant de 5 CRP (Centre Régional Psychotraumatisme) différents (Haut-de-France, Poitiers, Caen, Sud Nouvelle-Aquitaine, Centre Val de

Loire). Parmi eux, 12 ont déclaré pratiquer la thérapie d'exposition en imagination à l'évènement traumatique dans leur prise en charge du TSPT.

4.2.5 Analyses

Les analyses statistiques que nous présentons sont faites en Python avec les modules `seaborn` (© M. Waskom, 2021-2023) `Pingouin` (© R. Vallat, 2018-2023) et `pymer4` (© E. Jolly, 2017-2023).
Deux types d'analyses sont réalisés :

1. - des test T de Student ; pour chacun desquels nous présentons la valeur du t et sa p-value.
2. - des analyses de régression linéaire mixte ; pour chacune, nous présentons l'équation du modèle (dans sa syntaxe *R*) et la pertinence de chaque paramètre du modèle sera illustrée par les résultats du test T de Student de nullité statistique du coefficient associé à chaque paramètre (H0 : $\beta = 0$, contre H1 : $\beta \neq 0$) sous la forme de la valeur du coefficient β d'estimation du paramètre, la valeur du t et sa p-value.

4.3 Résultats

4.3.1 Les soignants sont capables de reconnaître quand le patient va mieux à partir de sa voix

Dans un premier temps, nous nous intéressons à la performance des soignants lorsqu'ils doivent comparer deux extraits non transformés S0 et Sf afin de savoir si les soignants sont capables de discriminer correctement lequel des deux correspond à la séance enregistrée quand le ou la patiente est guérie.
Pour cela, nous calculons les hits (essais pour lesquels la réponse est supérieure à 50, indiquant que le thérapeute a bien jugé le patient dans l'extrait Sf comme allant mieux que dans l'extrait S0). Puis nous comparons la moyenne des hits (M=0.613) au hasard par un test T de Student à une condition. Les résultats [$t(16) = 3.59$; $p = .0002$] montrent que les soignants évaluent bien les patients dans les extraits Sf guéris comme allant mieux que dans les extraits S0 de manière significative.

De plus, la tâche semble d'autant mieux réussie que les soignants pratiquent la thérapie d'exposition en imagination. En effet les résultats du T-test chez les thérapeutes ne la pratiquant pas [M= 0.57, $t(16) = 1.20$; $p = .12$] ne sont pas significatifs alors qu'ils le sont chez ceux qui l'utilisent [M= 0.63, $t(16) = 3.50$; $p = 0.0003$] - néanmoins la différence entre les deux groupes n'est pas significative (t=0.64 p=0.52) (Figure 4.4).

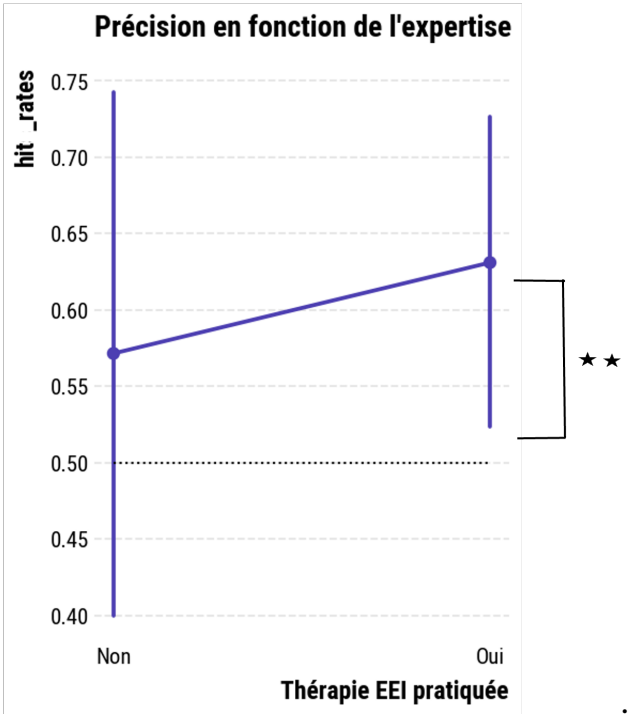


FIGURE 4.4 – Précision des réponses des soignants chez les experts et non experts de la thérapie d'exposition en imagination. Les astérisques indiquent des différences significatives et les barres d'erreurs correspondent à des intervalles de confiance de 95%. La ligne pointillée représente la performance au hasard

4.3.2 Les soignants se basent sur le pitch de la voix pour faire ce jugement

Maintenant que l'on sait que les soignants pratiquant la thérapie d'exposition savent estimer que le patient va mieux dans l'extrait Sf que dans l'extrait S0, nous examinons l'influence du pitch dans cette évaluation via l'analyse de l'effet des différentes transformations de pitch opérées sur l'évaluation des soignants. Pour cela nous utilisons le modèle (4.1) dans les deux types de transformations (baisse de pitch à

partir de S0 et hausse de pitch à partir de Sf) :

$$\text{response} \sim \text{percent\_transformed} + (1|\text{subject\_id}) \tag{4.1}$$

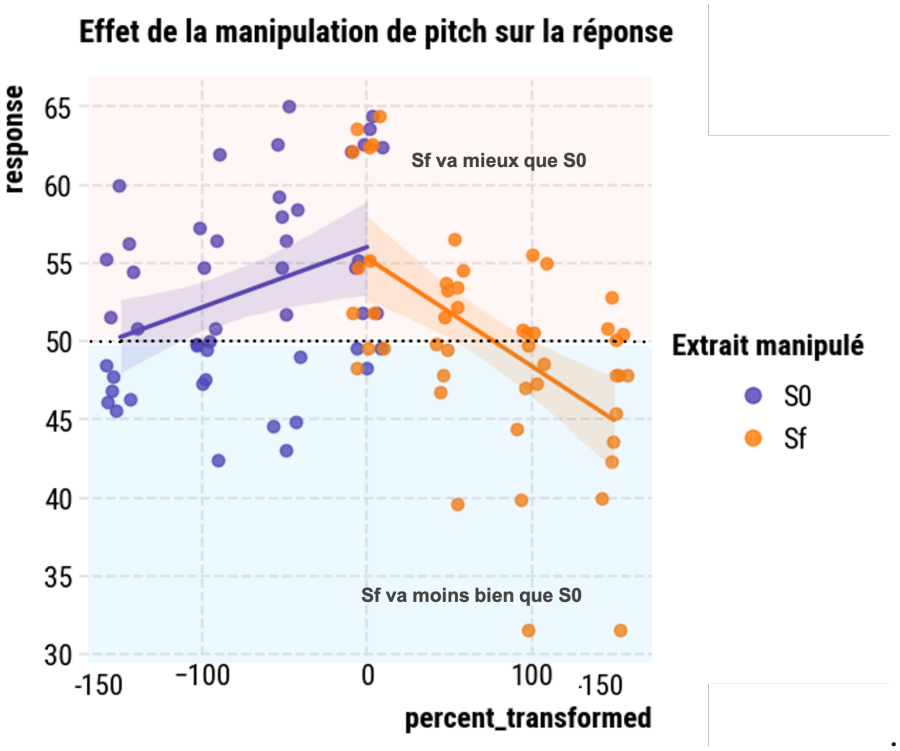


FIGURE 4.5 – La réponse supérieure à 50 signifie que le soignant indique que le patient va mieux dans l'extrait Sf par rapport à l'extrait S0. La variable percent_transformed décrit la manipulation faite sur l'extrait : soit en baissant le pitch de S0 [de -50 à -150](violet, à gauche), soit en augmentant le pitch de Sf [de 50 à 150](orange, à droite). Sf : session finale, en fin de thérapie, S0 : session initiale, en début de thérapie.

On observe que la valeur de la variable `response` diminue en fonction de la transformation, que ce soit pour les transformations de S0 ((β=0.038 95% CI [-0.004 – 0.081], t(16)=1.759, p=0.08) ou les transformations de Sf ((β= -0.069 [-0.112 – -0.026], t(16)=-3.137, p=0.002) ; 4.5). A gauche, la baisse de pitch de S0 de 0 à 150 amène progressivement la réponse des soignants à 50, càd la zone pour laquelle les soignants ne semblent pas entendre de différence dans l'état du patient entre les deux extraits présentés. A droite, la hausse de pitch sur Sf de 0 à 100% a le même effet sur la réponse des soignants qui avoisine la zone des 50. Par contre, pour la transformation qui consiste à augmenter le pitch de Sf de 150%, la réponse des soignants passe sous

la limite des 50, indiquant que le patient en Sf+150% est perçu comme allant moins bien qu'en S0.

4.3.3 L'effet de la manipulation de pitch ne s'observe que chez les soignants experts de la thérapie d'exposition en imagination

L'application du modèle (4.1) aux données des soignants ne pratiquant pas la thérapie d'exposition ne permet pas de mettre en évidence d'effet du pitch dans leurs jugements. Aucune relation linéaire significative n'est ainsi mise en évidence que ce soit pour les transformations opérées sur S0 ((β= -0.029 [-0.11 – 0.061], $t(16)$=-0.62, p=0.53) ou sur Sf ((β=-0.039 [-0.182 – 0.104], $t(16)$=-0.536 , p=0.593). L'interaction entre l'expertise et le degré de transformation, testée avec le modèle response $\sim$ percent_transformed $\times$ exposition en imagination + (1|subject_id) n'est pas significative (Figure 4.6).

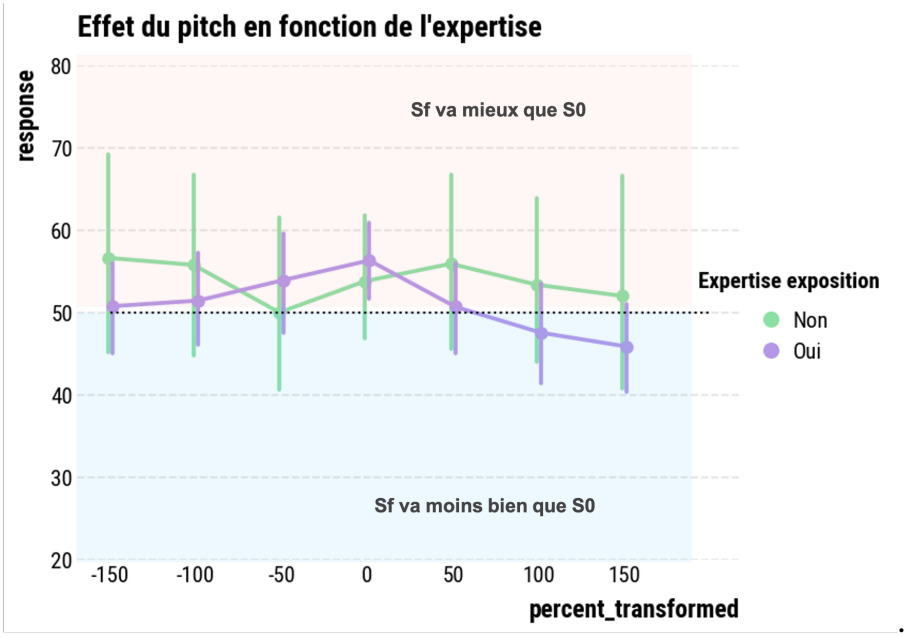


FIGURE 4.6 – Effet du pitch sur la réponse des soignants en fonction de la pratique ou non de la thérapie d'exposition en imagination. La réponse supérieure à 50 signifie que le soignant indique que le patient va mieux dans l'extrait Sf par rapport à l'extrait S0. La variable percent_transformed décrit la manipulation faite sur l'extrait : soit en baissant le pitch de S0 [de -50 à -150](à gauche), soit en augmentant le pitch de Sf [de 50 à 150](à droite). Sf : session finale, en fin de thérapie, S0 : session initiale, en début de thérapie.

4.3.4 L'application d'un filtre pitch sur les extraits malades ne permet plus de les différencier d'extraits guéris

Sachant que le pitch semble avoir un effet sur le jugement des soignants experts de la thérapie d'exposition, nous souhaitons observer plus en détail, chez ces soignants, l'effet de chaque type de filtre sur la performance à la tâche.
Pour cela, pour chaque niveau de transformation, nous réalisons un test T de student sur les hits en suivant la même méthode qu'en section 4.3.1. Nous cherchons ainsi à tester si, pour chaque degré de transformation de pitch appliquée, les deux extraits sont toujours jugés comme significativement différents et plus précisément si le patient est toujours perçu comme allant mieux en Sf.

Effet des filtres appliqués sur extraits de patients en début de thérapie (S0, symptomatologie TSPT active) :
Les résultats montrent que l'application des filtres -100% [$t(16) = 0.25$; $p = 0.80$, 95% CI [0.41, 0.62]] et -150% [$t(16) = 1.61$; $p = 0.136$, 95% CI [0.48, 0.61]] ne permet plus aux soignants de juger le patient comme allant mieux en Sf qu'en S0. Cet effet est également observé mais de manière moindre pour le filtre -50% [$t(16) = 1.82$; $p = 0.096$, 95% CI [0.48, 0.71]].

Effet des filtres appliqués sur extraits de patients en fin de thérapie (Sf, symptomatologie TSPT amendée) :
Les résultats montrent que l'application des filtres +50% [$t(16) = 1$; $p = 0.34$, 95% CI [0.46, 0.61]] et +100% [$t(16) =$ -0.76 ; $p = 0.46$, 95% CI [0.36, 0.57]] ne permet plus de juger le patient comme allant mieux en Sf qu'en S0. L'effet du filtre +150% [$t(16) =$ -3.44 ; $p = 0.0055$, 95% CI [0.32, 0.46]] se distingue des deux autres car il semble amener les soignants à percevoir que le patient va moins bien en Sf+150% qu'en S0.

4.4 Discussion

Cette étude réalisée sur un groupe de 17 soignants spécialisés dans la prise en charge du TSPT examine de manière inédite la perception que ces soignants font de l'état de santé psychologique de patients à partir de l'écoute de leur voix ainsi que l'influence d'un filtre de pitch sur cette perception. Pour cela, nous avons demandé à

des soignants de comparer des extraits de voix deux à deux, correspondant au même moment du script traumatique, enregistré en début et en fin de thérapie. Selon les conditions, l'un des deux extraits présentés avait subi une transformation de pitch (baisse pour les extraits initiaux et hausse pour les extraits finaux). Nous discutons ici les résultats de ce travail qui nous semblent apporter plusieurs contributions. De fait, ils souligneraient d'une part l'utilisation du pitch comme indice vocal dans l'évaluation de la psychopathologie d'un patient et d'autre part, via la manipulation causale de cet indice, ils éclaireraient les effets qu'un filtre vocal de pitch pourraient avoir sur les jugements des soignants.

Tout d'abord, il apparaît que les soignants sont effectivement capables d'évaluer l'état de guérison d'un patient pendant la thérapie d'exposition en imagination à l'ET à partir de sa seule voix. Cette compétence n'est attestée de manière statistiquement significative que parmi les soignants pratiquant la thérapie d'exposition (non retrouvée chez les autres soignants prenant en charge le TSPT par d'autres thérapies), ce qui suggère qu'elle est le fruit d'un apprentissage par l'expérience.

L'importance de l'expertise a été grandement soulignée au sein de la recherche sur le diagnostic médical et ces données mettent l'accent sur l'expérience plutôt que sur de meilleures habiletés générales de raisonnement pour expliquer l'acuité du diagnostic chez les praticiens experts(Bower, 1979 ; Elstein et al., 1978 ; Eva et al., 2007) .

Ici, sachant que notre cohorte est entièrement constituée de soignants experts du TSPT dont certains se distinguent par la pratique de cette forme spécifique de thérapie, la différence de performance s'expliquerait non pas par une formation médicale générique (*comment « sonne » la voix d'un patient TSPT*), mais par l'apprentissage de régularités au cours de la situation clinique précise que représente cette forme particulière de thérapie (*comment « sonne » la voix d'un patient souffrant de TSPT en cours d'exposition en imagination*).

Plus précisément, les résultats de notre étude suggèrent que les soignants experts utilisent le pitch pour faire cette inférence au sujet de l'état du patient. De fait, les différentes valeurs de pitch manipulées artificiellement ont un effet sur la réponse des soignants qui tend à être significatif pour les transformations opérées sur les extraits initiaux et est clairement significatif pour celles réalisées sur les extraits finaux. Aussi les données de cette étude suggèrent que les indices vocaux pourraient effectivement faire partie du modèle ou « script » (au sens de Bower (1979)) de la pathologie TSPT

auxquels ces soignants se réfèrent lors du processus de raisonnement clinique. Plus précisément, l'heuristique utilisée alors serait fondée sur la relation entre pitch et gravité des symptômes de TSPT que nous avons mise en évidence dans la cohorte TraumacoustiK, relation qui pourrait être implicitement apprise et relèverait donc de processus de raisonnement clinique non analytiques ((Epstein, 1999 ; Eva, 2005 ; Pelaccia et al., 2011). D'une certaine manière ces résultats nous invitent à penser que les participants soignants connaissaient déjà, implicitement le résultat de l'étude TraumacoustiK avant que nous la fassions. Nous posons que cet apprentissage serait fait de manière implicite car aucune mention n'est faite au sujet de la voix du patient lors de la formation dispensée aux soignants des différents CRP (formation dont nous avons d'ailleurs assuré une partie de l'élaboration) . En plus de nous renseigner au sujet des processus cognitifs à l'oeuvre pendant le raisonnement diagnostique, ces résultats ont une portée clinique évidente et vont dans le sens de récents travaux dans le domaine de l'enseignement clinique qui suggèrent que les enseignants devraient souligner l'importance des deux formes de raisonnement, consciente/contrôlée d'une part et inconsciente/automatique de l'autre (Eva et al., 2007).

Cette heuristique mise en évidence, nous avons observé de manière plus détaillée l'effet de chaque type de filtre appliqué aux extraits de voix de patients. Nos résultats indiquent que l'abaissement de pitch de 0 à -150% appliqué aux extraits de voix enregistrés en début de thérapie amène progressivement la réponse des soignants à 50, càd à vraisemblablement ne plus entendre de différence dans l'état du patient entre ces extraits transformés et ceux correspondant réellement aux séances pour lesquelles le patient est considéré guéri. La diminution artificielle de pitch sur les extraits initiaux conduirait donc à les approcher perceptivement des extraits « guéris ». Parallèlement, la hausse de pitch de 0 à +100% appliquée sur les extraits enregistrés une fois le patient guéri, a un effet similaire sur la réponse des soignants dans le sens où elle approcherait perceptivement les extraits transformés des extraits naturels enregistrés quand le TSPT est pleinement actif.

Au delà de 100% de transformation, l'effet sur le jugement des soignants semble différer entre les deux types de transformations opérées (augmentation ou baisse). De fait, seule l'augmentation du pitch à 150% appliquée sur les extraits de fin de thérapie semble amener les soignants à une inversion de l'évaluation de l'état psychologique du patient entre S0 et Sf : les soignants jugeant que le patient va plus mal dans l'extrait

de fin de thérapie transformé à 150% que dans l'extrait naturel de début de prise en charge. Cet effet de la transformation nous semble d'autant plus intéressant qu'avant la transformation de pitch, ces mêmes soignants avaient un jugement purement inverse. A notre sens, ce résultat nous permet de répondre à la question de la part du pitch dans l'évaluation par la voix de l'état psychique du patient. En effet, il semblerait que le pitch de l'extrait transformé de fin de thérapie - alors plus élevé de 50% par rapport à celui de l'extrait de début de prise en charge - prime sur tous les autres indices acoustiques portés par la voix du patient souffrant activement de TSPT et ait ainsi le potentiel d'embarquer le jugement du soignant. Le fait que les effets ne soient pas complètement symétriques entre les deux types de transformation pourrait s'expliquer par une différence dans les indices acoustiques autres que le pitch portés par les extraits naturels sur lesquels nous appliquons les transformations de pitch.

Limites de l'étude :

Comme précisé en section Méthode, les paires de stimuli que nous avons constituées pour cette étude sont composées d'extraits qui correspondent au même moment du script traumatique enregistré en début et fin de thérapie. Or, s'agissant de matériel naturel, provenant de situations de thérapie réelles, et malgré un travail de sélection minutieux parmi le large corpus d'enregistrements dont nous disposions, les extraits vocaux ne sont donc pas littéralement identiques entre ces deux séances. Les limites que nous avons dégagées au chapitre précédent s'appliquent par conséquent également ici : d'autres indices acoustiques que nous n'avons pas exploré pourraient être recherchés et contribuer au jugement des soignants. Cependant, même sans avoir connaissance des autres paramètres acoustiques qui pourraient entrer en ligne de compte, nos résultats montrent que le pitch est un paramètre *suffisant* pour guider le jugement des soignants. Le contenu linguistique du discours n'a pas été analysé non plus, nous ne pouvons donc pas soutenir avec des données chiffrées que celui-ci n'influence pas le jugement des participants. Tout en reconnaissant ce biais possible, nous tenons à préciser que si influence du contenu du discours il y avait, celle-ci pourrait ne pas aller dans le sens des résultats que nous avons observés. En effet, à mesure des séances, l'amnésie traumatique se levant, le script de l'ET s'enrichit de plus en plus de détails à haute signification émotionnelle, une analyse quantitative de la sémantique des extraits pourrait conclure que les extraits des séances finales sont plus chargés émotionnellement qu'en début de thérapie.

Pour pallier ces différences de base entre les extraits comparés, nous aurions pu choisir de comparer deux extraits identiques pour lesquels seul le pitch changeait. Cependant, la comparaison d'un enregistrement original avec plusieurs de ses transformations rend particulièrement saillant le fait que toutes les transformations reproduisent l'intonation et le rythme de l'original de façon exactement identique, ce qui aurait pu créer un effet de demande important dans cette version alternative de l'expérience. Néanmoins, même en ne choisissant pas le design expérimental alternatif que nous venons de décrire, le fait que nous ne présentions qu'un nombre limité d'extraits dans notre expérience ne nous permet pas complètement d'écarter l'éventualité d'un effet de demande. Sachant que nous présentons aux mêmes participants plusieurs versions modifiées des mêmes extraits, ils pourraient percevoir qu'il s'agit des mêmes paires de sons, entendre qu'il y a quelque chose de gradé qui change à chaque fois (le pitch), et accorder leur réponse à cela. Conscients de ce possible biais au moment de l'élaboration du protocole, nous avons pris différentes précautions afin de limiter ce risque. Tout d'abord, nous avons pris soin d'exposer aux participants une cover-story afin de les détourner de l'objectif réel de l'étude, comme il est classique de faire dans les expériences de psychologie cognitive : *« Certaines paires vont peut-être vous sembler similaires, mais en réalité elles sont toutes différentes car correspondent à des enregistrements réalisés dans des conditions légèrement différentes. Il est donc très important que vous évaluiez chaque paire indépendamment les unes des autres et de ne pas vous baser sur vos réponses précédentes afin de pouvoir mesurer l'effet potentiel de ces conditions d'enregistrement sur votre perception clinique. »*. Ensuite, afin de limiter le plus possible le risque de reconnaître la transformation, nous avons adopté un design expérimental qui éloigne le plus possible la présentation des paires correspondant à des versions transformées du même extrait, dans des blocs différents. Enfin, les transformations opérées sur le pitch sont suffisamment faibles (de l'ordre d'un demi-ton) pour rester subtiles et donc difficiles à remarquer et à grader les unes par rapport aux autres. Le fait que nous ne retrouvions pas l'effet de manipulation de pitch sur les soignants non experts de la thérapie d'exposition en imagination tend à écarter l'hypothèse d'un effet de demande et va plutôt dans le sens d'une règle observée (par l'expérience) et intégrée au modèle cognitif des soignants experts.

4.4.1 Conclusion

En conclusion, il apparaît que les soignants experts de la thérapie d'exposition en imagination soient capables de juger de l'état de guérison de patients à partir de leur seule voix.

Cette compétence que nous mettons ici en évidence reposerait sur une règle apprise implicitement à l'écoute des séances d'exposition en imagination. Dès lors, si l'on applique les principes de l'inférence bayésienne décrits en introduction de ce manuscrit, la fréquence des occurrences de ces régularités pendant la pratique des experts amènerait à accorder à cette croyance un poids notable au sein du modèle cognitif des experts . De fait, les soignants experts semblent lui accorder un crédit substantiel, cette heuristique guidant leur jugement au point de les induire en erreur lorsqu'on manipule ici les extraits vocaux par un filtre de pitch. Ils jugeraient alors qu'un patient est malade alors qu'il ne l'est pas. Avant cela, les travaux de Elstein (1999) utilisant des modèles bayésiens pour étudier la prise de décision médicale avaient ainsi déjà mis en évidence le fait que les heuristiques comptent parmi les facteurs cognitifs (avec les biais) qui produisent des erreurs dans le raisonnement clinique.

Face à de tels résultats, ce travail permettrait à notre sens d'apporter un premier argument en faveur de notre hypothèse élaborée en introduction qui propose l'influence du filtre vocal sur les processus d'inférence perceptive ainsi qu'un début de contribution à la question de savoir ce que l'objet filtre pourrait faire à nos cognitions sociales fines en situation d'interaction. Cette dernière est observée dans le cadre de la prise en charge thérapeutique mais nous pouvons tout naturellement imaginer que les effets d'un filtre de pitch ne se limiteraient pas à celui-ci.

Les résultats de l'extension de l'étude TraumacoustiK (Section 3.5) nous ont permis de dégager un lien entre pitch et rythme cardiaque. En plus de nous informer au sujet de l'état d'un patient, le pitch pourrait-il porter de l'information au sujet d'un autre état caché de notre interlocuteur, à savoir son activité cardiaque ? Quelle serait alors l'incidence d'un filtre de pitch sur ces inférences ? Le chapitre 5 propose d'adresser cette question en deux temps. Tout d'abord, nous examinerons plus en détail le lien pitch-RC chez les volontaires sains. Nous choisissons de travailler alors dans cette population car il nous semble qu'elle permettrait une observation moins biaisée (cf perturbations du

SNV chez les patients souffrants de TSPT) de la relation pitch - RC et ainsi un meilleur contrôle expérimental. Puis, nous testerons, toujours au sein d'une cohorte de volontaires sains, la capacité des individus à percevoir le RC d'un locuteur au moyen de sa seule voix ainsi que l'influence potentielle du pitch et sa manipulation artificielle sur ce jugement.

Déclaration de contribution

Nadia Guerouaou : Conception de l'étude, Analyse des données statistiques, Rédaction - préparation du projet original.
Coralie Vincent (Ingénieure de recherche) : Design expérimental et Implémentation de l'étude en ligne, Rédaction - révision du projet.
Frédérique Warembourg, Wissam El Hage, Mélanie Voyer, Amaury Mengin, Chantal Bergey (chefs de service de CRP) : recrutement des personnels soignants
JJ Aucouturier (co-encadrant) : Conception de l'étude, Création des stimuli acoustiques, Analyse des données statistiques, Rédaction - préparation du projet original.
Guillaume Vaiva (co-encadrant) : Conception de l'étude, Rédaction - révision du projet.

5. Étude de la relation pitch – rythme cardiaque

De la production à la perception

5.1 Volet production : Expérience de tilt test

La mise en évidence au sein de l'extension de notre étude TraumacoustiK (Section : 3.5) d'un lien entre les valeurs de pitch et le rythme cardiaque (RC) vient renforcer un faisceau assez faible d'évidences corrélationnelles dans la littérature qui mettent en évidence l'existence d'une covariation entre les signaux vocaux et cardiaques.

Plusieurs études suggèrent en effet que des circuits nerveux communs pourraient entraîner des covariations entre l'état physiologique d'un locuteur et les paramètres acoustiques retrouvées dans la voix Stewart et al. (2013). Citons par exemple les mêmes structures au niveau du tronc cérébral qui sont impliquées à la fois dans la régulation du rythme cardiaque via la branche myélinisée du nerf vague et les vocalisations via les muscles laryngés et pharyngés Ayres and Gabbott (2002). Ainsi les afférences vagales vers ces derniers muscles influençant les caractéristique prosodiques de la vocalisation pourrait refléter l'influence vagale au niveau cardiaque.

S'intéressant aux causes des perturbations naturelles de la fréquence fondamentale dans la voix, le travail d'Orlikoff and Baken (1989) a pu mettre en évidence que ces variations longtemps considérées comme complètement aléatoires pouvaient en réalité être en partie expliquées par une influence modeste mais constante du système cardiovasculaire rendant compte d'environ 0,5 % à 20 % de la perturbation absolue de la F0 ("jitter") mesurée au cours d'une phonation soutenue.

Ces données corrélationnelles ont inspiré par ailleurs des travaux en ingénierie biomédicale visant à créer des outils d'apprentissage machine capable d'extraire le RC d'un individu à partir d'enregistrements vocaux. Plusieurs algorithmes permettant d'estimer le RC à partir de multiples caractéristiques vocales ont ainsi été développés et montrent la faisabilité de l'estimation automatique de la fréquence cardiaque à partir de la voix humaine, en particulier à partir des voyelles soutenues(Sakai, 2015 ; Schuller et al., 2014 ; Usman et al., 2021), (dans l'habituelle mesure où ces apprentissages parviennent à se généraliser en dehors d'un corpus d'apprentissage donné). Cependant, compte tenu des objectifs initiaux de ces études, elles ne nous informent que peu au sujet de la nature de la relation - linéaire, monotone ou plus complexe - entre le pitch et le RC.

De fait, « *bien que la littérature existante suggère que le processus de production de la parole est affecté par des changements physiologiques chez les individus, l'effet de ces changements physiologiques sur les paramètres réels de la parole nécessite une étude approfondie* » (Trouvain and Truong, 2015). En effet si le lien de corrélation entre pitch et RC que nous avons mis en évidence dans l'étude TraumacoustiK semble étayé dans la littérature, la causalité de cette relation reste encore à être interrogée.
Pour ce faire, nous proposons de concevoir un protocole expérimental au cours duquel le RC des participants serait directement manipulé en vue d'observer les effets provoqués par cette manipulation sur leur voix, et en particulièrement sur le pitch. Cette manipulation vise à observer si les effets potentiels sur le signal vocal sont alignés avec la relation vont dans le sens de ceux attendus au regard de la covariation mise en évidence dans le chapitre 3.
Plusieurs paradigmes expérimentaux permettent de créer de manière non invasive des modifications du RC en laboratoire, comme la course sur tapis. Cependant, comme nous souhaitons que les modifications observées sur la voix traduisent le plus directement possible la variation du RC, la manipulation idéale doit avoir le moins d'effets possibles (ex. respiratoires) ayant eux-aussi une influence possiblement confondante sur la voix. C'est le cas de l'activité physique dont on sait notamment qu'elle entraîne des modifications des fréquences centrales des formants, du placement des pauses respiratoires et du pitch (Primov-Fever et al., 2014). Nombre de ces changements sont dus à la compétition interne du sujet entre la parole et la respiration pendant l'exécution d'une tâche physique, ce qui a un impact corollaire sur le contrôle musculaire et

le flux d'air au niveau sous glottal ainsi que dans la structure articulatoire du tractus vocal (Godin and Hansen, 2015).

Afin de créer des variations de RC influençant le moins possible l'activité respiratoire nous avons donc établi un paradigme expérimental inspiré du protocole de *tilt-test* (ou « test d'inclinaison »), un examen clinique qui permet de reproduire les perturbations vago-sympathiques observées au cours des syncopes vaso-vagales. Ce protocole consiste, tout d'abord, à allonger le participant pendant une durée d'environ 15 minutes, puis de le relever rapidement. Après une station allongée, la mise en position debout entraîne une redistribution de 500 à 1 000 mL de sang de la partie supérieure du corps vers les membres inférieurs. Cette diminution du retour veineux entraîne une diminution du diamètre du ventricule gauche, ce qui provoque une accélération de la fréquence cardiaque (Aponte-Becerra and Novak, 2021). Le retour à un niveau de base pour cette dernière est attendue environ 5min après le changement de station. Dans notre protocole, afin d'observer de potentielles perturbations vocales causées par une modification du RC, les participants, après une période de repos allongés passaient à la station debout et étaient invités à vocaliser (tenir la voyelle /a/). Les productions vocales étaient alors enregistrées ainsi que, de façon synchronisée, l'activité cardiaque du locuteur au moyen d'une méthode de photopléthysmographie (PPG). Il s'agit d'une technique qui utilise une source lumineuse et un récepteur pour quantifier la lumière absorbée ou réfléchie par la peau humaine. De fait, le volume de sang dans nos tissus influe sur la quantité de lumière que la peau absorbe ou réfléchit. Le PPG offre la possibilité d'estimer le rythme cardiaque de manière non invasive. D'autres mesures (enregistrements vidéos et activité respiratoire) ont également été faites lors de ce protocole mais nous faisons le choix de ne présenter ici que les données s'inscrivant directement dans le cadre de notre problématique, à savoir les données vocales et cardiaques.

5.1.1 Matériel et méthodes :

5.1.1.1 Participants

Après signature de formulaire de consentement de l'étude, 27 participants, âgés de 18 à 30 ans (femmes :14) ont participé à l'expérience. Ceux -ci ont été recrutés en suivant les critères suivants :

Critères d'inclusion : être capable de lire un petit texte à haute voix en français, être capable de se coucher et de se lever sans faire de malaise, français langue maternelle.

Critères d'exclusion : maladie cardiovasculaire, pathologie de la voix (bégaiement, etc.), traitement pharmacologique altérant la fonction cardiaque (ex. bêtabloquant).

Les participants étaient identifiés par leur numéro d'inclusion dans l'étude, nous ne conservons aucune information permettant une identification directe. Leur participation était indemnisée à hauteur de 12€.

5.1.1.2 Ethique

L'expérience s'est déroulée en juillet 2022 au Centre Multidisciplinaire des Sciences Comportementales INSEAD/Sorbonne Université, et a été autorisée par l'IRB INSEAD (décision du 8 Avril 2022, numéro de référence 2022-24).

5.1.1.3 Procédure

L'expérience débute par une phase de repos allongée de 15 minutes. Ensuite, les participants se levaient et commençaient la phase de vocalisation. Lors de celle-ci, le participant vocalisait un /a/ à voix haute pendant un maximum de 10 secondes (avec pour consigne de vocaliser de manière la plus naturelle possible, sans forcer sur la voix et de s'arrêter avant de s'essouffler ou de s'égosiller). Ceci est répété 15 fois. Chaque essai dure 10 sec et est suivi de 5 secondes de pause. Le début des essais est déclenché par l'expérimentateur qui appuie sur la barre d'espace de l'ordinateur déclenchant l'apparition à l'écran de la consigne "Vocalisez a". Les 15 essais sont réalisés successivement pendant une période de 5 mn.
L'expérience était précédée d'une phase d'installation de 15mn. Avant de commencer l'enregistrement, nous expliquions l'expérience au participant et nous placions les capteurs. Une fois l'enregistrement terminé, nous prenions un temps pour debriefer avec le participant, en leur expliquant le but de l'étude et en répondant à leur diverses questions éventuelles. La durée totale de l'expérience était d'une heure et demie.
L'expérience a été codée en Python et est disponible sur github.com/neuro-team-femto/dataset_voix_coeur.

5.1.1.4 Mesures

Enregistrement audio : les vocalisations des participants étaient enregistrées au moyen d'un micro voix (de type Neumann TLM102) placé à 50cm des participants, relié à une carte d'acquisition audio RME Fireface.
Enregistrement physiologique : Les données de rythme cardiaque ont été acquises à l'aide d'un capteur de photopléthysmographie (PPG) fixé à l'index du participant et connecté à un amplificateur EEG Actichamp (Brain Systems, Allemagne) en faisant l'acquisition à une fréquence d'échantillonnage de 1000Hz. Le signal était enregistré sur un ordinateur portable à l'aide du logiciel BrainRecorder. La synchronisation entre les différentes mesures était assurée par un boîtier Cedrus StimTracker.

5.1.1.5 Traitement des signaux

Pré traitement des données : Les données enregistrées sont découpées en trois phases : les phases de repos, de silence et de vocalisation. Nous traitons ici les données correspondant aux phases de vocalisation.
Données vocales :

— Réduction de bruit : A l'analyse, il apparaît que la prise de son pour les 9 premiers sujets a été perturbée par un bruit basse-fréquence, possiblement de l'enregistrement de la ventilation de la salle. Pour palier à cela, nous appliquons sur l'ensemble des enregistrements une réduction de bruit en utilisant l'algorithme « Noise Reduction » du logiciel Audacity (© 1999-2021 Audacity Team) avec un gain de 28dB, une sensitivité de 6 et un lissage en fréquence sur 3 bandes.

— Extraction du pitch : Afin d'obtenir les valeurs moyennes de pitch, nous utilisons ensuite le logiciel Praat (Boersma P., 2007), pour extraire le paramètre *mean pitch* sur une fenêtre courante de 10ms, puis en moyennant ces valeurs sur une fenêtre équivalente à l'ensemble de l'extrait.

Données PPG : extraction du rythme cardiaque :
Les données PPG sont analysées en Python (vers. 3.8.10) en utilisant la procédure par défaut du module heartpy (vers. 1.2.7 (van Gent et al., 2019)) :

— Tout d'abord, nous appliquons un filtre passe-bande (entre 0.7Hz et 3.5Hz, d'ordre 3).

— A chaque point temporel, nous définissons un seuil pour détecter les pics élevés en calculant la moyenne d'une fenêtre de 750 ms centrée sur ce point (technique du « moving average »).
— Toutes les données supérieures à ce seuil sont définies comme des pics ; pour chaque pic, nous considérons la position du maximum comme la position de la systole (ou "pic R").
— Les intervalles RR (intervalle de temps entre deux pics R) sont calculés entre les pics R successifs, puis convertis en minutes, moyennés et inversés pour obtenir une valeur moyenne de BPM pour la séquence analysée.
— Si le BPM obtenu est inférieur à 40 ou supérieur à 180, ou si l'écart-type du RR est supérieur à 0.1 ms, cette procédure est répétée avec un seuil progressivement plus élevé, augmenté par paliers de 5 % jusqu'à ce que ces conditions soient remplies.
— Enfin, les pics R sont supprimés si leurs intervalles RR précédents sont des valeurs aberrantes définies comme +/- 30 % de l'intervalle RR moyen. La valeur finale du BPM est calculée comme ci-dessus, en utilisant cette liste finale de pics.

Ces analyses nous permettent d'obtenir une valeur moyenne de pitch et de bpm pour chacun de nos 15 extraits.

5.1.1.6 Analyses statistiques

Compte tenu d'un problème technique concernant l'acquisition des données physiologiques pour le participant 4, seuls 26 participants (F=13) ont été inclus dans notre analyse. Les analyses statistiques que nous présentons sont faites en Python, avec les modules `seaborn` (© M. Waskom, 2021-2023) `pingouin` (© R. Vallat, 2018-2023) et `pymer` (© E. Jolly, 2017-2023). Pour les analyses de régression, `pymer` utilise le module `lmer` qui estime le modèle testé à l'aide de la méthode de vraisemblance restreinte ou résiduelle (acronyme anglais REML) et les p values ainsi que les degrés de libertés (DF) sont estimés par l'approximation de Satterthwaite. Qu'il s'agisse de modèle de régression linéaire simple ou mixte, nous présentons l'équation du modèle (dans sa syntaxe *R*). La pertinence de chaque paramètre du modèle sera illustrée par les résultats du test T de Student de nullité statistique du coefficient associé à chaque

paramètre (H0 : $\beta = 0$, contre H1 : $\beta \neq 0$) sous la forme de la valeur du coefficient β d'estimation du paramètre, la valeur du t et sa p-value.

5.1.2 Résultats

Considérant les différences physiologiques entre les individus de sexes masculin et féminin dans les mesures de pitch, et les données de la littérature en faveur d'un possible effet du sexe sur l'activité cardiaque (Umetani et al., 1998 ; Wadhwa et al., 2008) les figures illustrant nos analyses présenteront les données en distinguant les participants de sexe masculin (M) et féminin (F).

5.1.2.1 Le RC tend à augmenter au cours des essais successifs

Afin d'observer si notre paradigme expérimental crée effectivement une variation de RC dans le sens recherché, nous avons dans un premier temps analysé le RC des participants et sa variation au cours des différents essais. En moyenne le bpm pendant les phases de vocalisation était de M(F)= 85.5 (SD=4.80) pour les participantes et M(M)= 94.8 (SD= 7.99) pour les participants.
Nous testons la relation entre le bpm et les essais successifs via le modèle linéaire suivant :

$$\text{bpm} \sim \text{trial\_number} + \text{sex} \tag{5.1}$$

Les résultats de l'estimation montrent une tendance non significative de l'effet de la variable trials number [$t(27)$=1.748, p= 0.092]. La valeur du coefficient beta [β=0.160] indique que le bpm tend à augmenter au fur et à mesure des essais. Au vu des résultats de notre modèle de régression du bpm moyen en fonction des essais, il semblerait que notre protocole tende bien à créer une variabilité dans le rythme cardiaque des participants au cours des essais successifs. La valeur du R^2=0.856 (ajusté pour facteurs multiples) tendant vers l'unité nous conforte sur la pertinence de ce modèle (Figure 5.1).

5.1.2.2 Le pitch moyen augmente significativement au cours des essais successifs :

Sur toute la période de vocalisation, le pitch moyen chez les participantes est de 253.1 Hz (SD=10.2) et 143.7 Hz (SD=7.8) chez les participants. Nous testons la relation entre le pitch moyen et les essais successifs via le modèle linéaire mixte suivant :

$$\text{mean\_pitch} \sim \text{trial\_number} + (1|\text{participant\_id}) \tag{5.2}$$

Le résultat du modèle de régression du pitch moyen en fonction des essais [β=1.027, $t(377)$= 8.586 , p= 0.0 ***] confirme l'augmentation significative du pitch moyen au fur et à mesure des essais que nous observons dans la Figure 5.1.

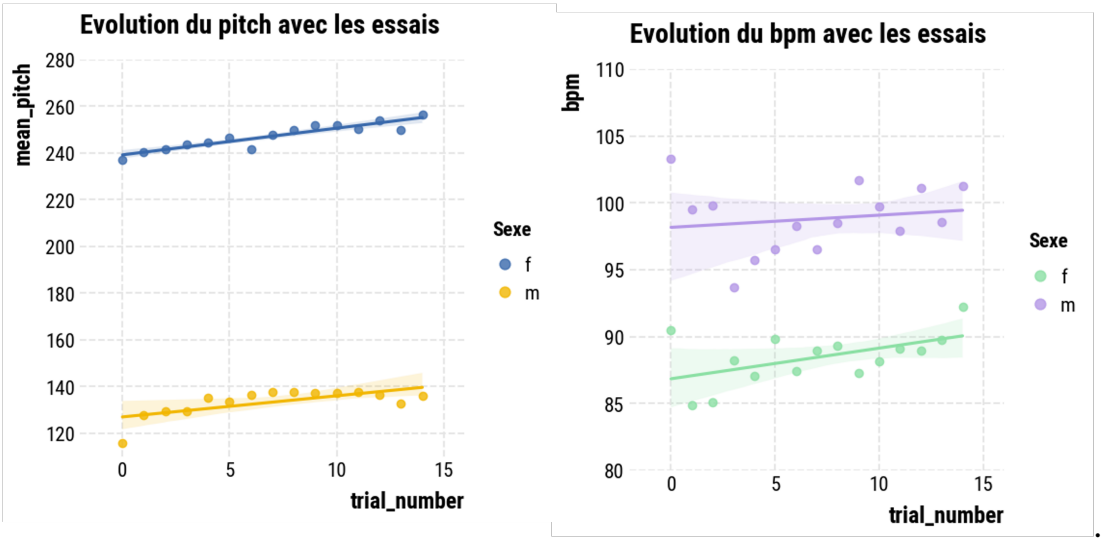


FIGURE 5.1 – Par souci de clarté de la visualisation, les deux graphiques illustrent la moyenne des valeurs de pitch à gauche et de RC à droite pour chaque essai, contrairement aux statistiques que nous décrivons dans le texte qui prennent en compte tous les essais individuels.

5.1.2.3 Mise en évidence d'une relation pitch- RC

La relation entre le pitch moyen et le RC au fur et à mesure des essais est testée par un modèle linéaire multiple selon l'équation suivante :

$$\text{mean\_pitch} \sim \text{bpm} + \text{trial\_number} + (1|\text{participant\_id}) \tag{5.3}$$

Nos résultats indiquent que les 2 facteurs ont une influence significative sur le pitch. D'une part, le trial_number (β=+0.70 [0.35 – 1.05], t(347.01)=3.93, p =0.00 ***) et de l'autre le bpm (β=+0.44 [0.24 – 0.63], t(355.67)=4.39 , p =0.00 ***) ont une influence positive sur le comportement du pitch (Figure 5.2.

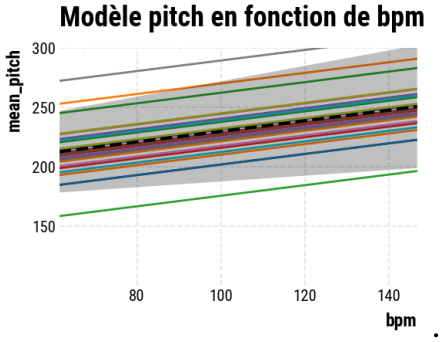


FIGURE 5.2 – Illustration des paramètres de la variable `bpm` (pente et ordonnée à l'origine) du modèle (5.3) affichés sous la forme de droites parallèles. Chaque couleur représente un participant.

5.1.3 Conclusion

Les données de notre étude Traumacoustik d'une part et celles dont font état la littérature de l'autre suggèrent une covariation entre signaux vocaux et cardiaques, possiblement sous-tendue par le partage de circuits neuronaux communs. Cependant, ces travaux se limitent à des données de nature corrélationnelle. Il nous a donc semblé nécessaire de tester expérimentalement la causalité de ce lien. Pour ce faire, nous avons conçu un protocole expérimental, inspiré de l'examen clinique de tilt-test, en vue de manipuler causalement (le plus spécifiquement possible) le RC cardiaque de participants afin d'observer les conséquences de la dite manipulation sur le pitch enregistré lors de la production de la voyelle /a/ soutenue.

Les résultats de cette étude nous ont permis d'attester de l'efficacité de ce nouveau paradigme expérimental pour créer des variations contrôlées du RC chez les participants. De fait, nous avons pu mettre en évidence une variation, bien que légère (β=+0.16), dans le RC des participants au cours des 15 essais alors même qu'ils étaient maintenus en dans une station debout immobile.

Ceci nous a permis d'observer le résultat principal de cette étude, à savoir une relation linéaire positive entre le RC des participants et les valeurs de pitch relevées lors de leurs vocalisations. Pour autant, il convient de souligner que la relation que nous trouvons ici entre les mesures de pitch et de bpm est de sens inverse par rapport à celle que nous avions mise en évidence au sein de notre corpus TraumacoustiK : extension (Section : 3.5.2.4. Aussi, nous avions alors discuté le fait que l'augmentation du pitch associée à une baisse du RC chez les patients ne soit pas retrouvée dans la littérature (Usman et al., 2021) en faisant l'hypothèse que le sens de la relation observée au sein de cette population souffrant de TSPT puisse être lié à une perturbation de la balance sympathique parasympathique décrite chez ces patients (Thayer et al., 2012). Dans la présente étude réalisée auprès de volontaires sains, le sens de la relation *pitch* $\leftrightarrow$ *bpm* suit les données de la littérature et pourrait refléter un fonctionnement dit physiologique. D'autre part, la différence entre nos deux résultats pourrait également s'expliquer par le fait que nous ayons ici mis en oeuvre une manipulation causale d'une de nos deux variables d'intérêt, passant alors de l'observation de données issues de corrélations à des données dont la preuve est plus forte au regard de la méthode expérimentale scientifique Casadevall and Fang (2008). Ainsi, comme le dit Raymond Aron dans sa préface de l'ouvrage de Weber and Colliot-Thélène (2003) intitulé Le Savant et le Politique, « la vocation de la science est inconditionnellement la vérité. Et cette quête de vérité ne saurait se satisfaire de simples corrélations, fussent-elles souvent vérifiées ; les relations causales, les mécanismes explicatifs et, surtout, les preuves demeurent essentielles à la compréhension ; à défaut, nous ne disposerons que de conjectures ».

Déclaration de contribution

Nadia Guerouaou : Conception de l'étude, Collecte des données, Analyse des données - statistiques, Rédaction - préparation du projet original.
Paul Maublanc (ingénieur de recherche) : Analyse des données - acoustiques et cardiaques
Matthieu Fraticelli : Design de la procédure d'enregistrement, Collecte des données
JJ Aucouturier (co-encadrant) : Conception de l'étude, Analyse des données - acoustique, Analyse des données - statistiques, Rédaction - préparation du projet original.
Guillaume Vaiva (co-encadrant) : Conception de l'étude, Rédaction - révision du projet.

5.2 Volet perception : Étude de l'inférence du RC dans la voix

Depuis les travaux récents de Galvez-Pol et al. (2022) et leur réplication par Arslanova et al. (2022), nous savons que les individus sont capables de détecter le rythme cardiaque chez autrui simplement en regardant leur visage présenté en vidéo. Pour cela, les auteurs ont développé une nouvelle tâche de choix forcé à deux intervalles (2I-1AFC) dans laquelle les participants regardaient des vidéos montrant deux locuteurs côte à côte et au centre un carré clignotant au rythme des battements de cœur de l'un des deux (Figure 5.2). La tâche des participants consistait à choisir à qui appartenait le battement cardiaque présenté au centre.

Les résultats de leurs expériences ont montré que les participants sont capables de déterminer le propriétaire le plus probable d'une séquence donnée de battements de cœur, et ce de manière nettement supérieure au hasard (précision interindividuelle allant de 45 % à 74 %, n=24 chez Galvez-Pol et al. (2022) et 40 % à 80 %, n=142 chez Arslanova et al. (2022)). Ces travaux constituent une première étape dans l'étude de la faculté à inférer le RC de nos interlocuteurs en observant leur visage. Cependant, les facteurs précis qui sous-tendent cette capacité doivent encore être établis.

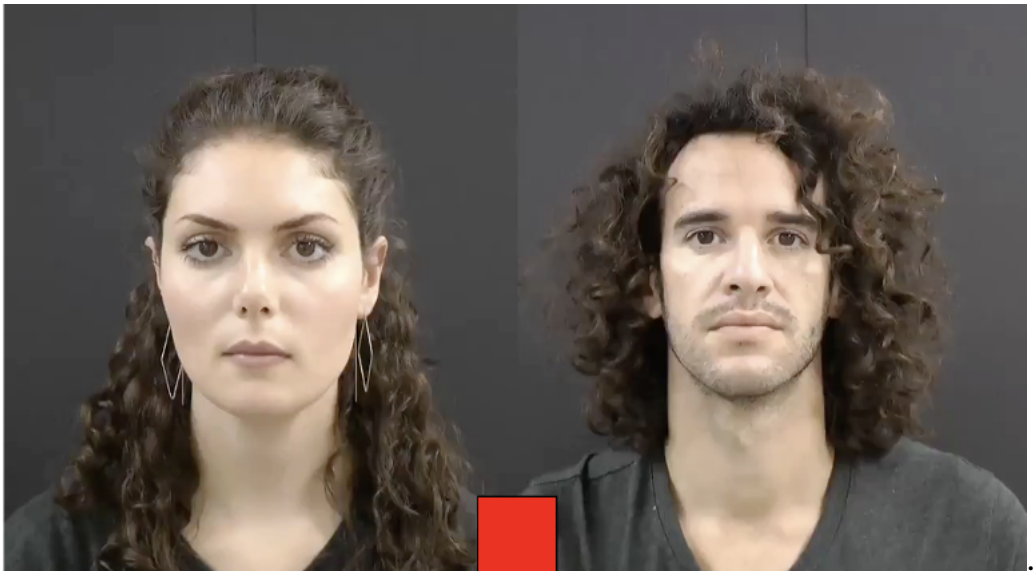


FIGURE 5.3 – Illustration de la tâche "The Other's Hear Rate", photo de la vidéo disponible via https://osf.io/6yt92

Nous proposons de penser ces résultats à l'intérieur du cadre de la théorie du codage prédictif qui stipule que nous appliquons des inférences ~~non conscientes~~ à tous les domaines de la perception. Il apparaît alors tout à fait probable, sachant l'évidence du lien entre activité cardiaque et état émotionnel, que ces informations au sujet du RC de notre interlocuteur, possiblement codées de manière non consciente, participent de nos inférences au sujet de l'état émotionnel de notre interlocuteur.

Dans cette thèse, nous interrogeons l'effet que des transformations vocales comme un « filtre de pitch » peuvent avoir sur notre cognition sociale fine. Cette question se décline en plusieurs : la première est celle de savoir si la voix véhicule de l'information au sujet du rythme cardiaque du locuteur que les individus seraient capables de percevoir (extension de l'étude de Galvez-Pol et al. (2022) à la voix). Puis, nous interrogeons la possibilité que cette information puisse être extraite d'un paramètre acoustique particulier, le pitch qui nous intéresse spécifiquement dans cette thèse et l'utilisation d'une heuristique pitch-rythme cardiaque fondée sur les régularités observées dans notre environnement. Ainsi, nous savons de par les données de notre précédente expérience tilt-test, qu'il existe une relation causale positive entre RC et pitch. Nos modèles génératifs du monde étant élaborés à partir des régularités observées au sein de notre environnement au cours de nos expériences passées, nous posons la question

de l'utilisation de cet indice vocal spécifique pour guider nos inférences au sujet de l'activité cardiaque de notre interlocuteur. La pertinence de ce paramètre acoustique nous est également réaffirmée par l'importance du pitch comme indice utilisé pour inférer les émotions dans la voix d'autrui (Juslin and Laukka, 2003). Pour cela, nous observerons l'effet de l'application d'un filtre pitch sur l'attribution d'une pulsation cardiaque à un extrait vocal dont toutes autres caractéristiques acoustiques restent inchangées. Nous espérons de par cette méthodologie pouvoir à la fois adresser la question biologique du rôle du pitch dans la perception du RC et celle, applicative, du potentiel de manipulation de l'objet filtre quant à la perception du signal cardiaque, dont on connaît l'importance en tant que témoin des états émotionnels de nos interlocuteurs (Beauchaine, 2015).

5.2.1 Hypothèses

Nous avons donc conçu une adaptation du paradigme de Galvez-Pol et al. (2022) afin de tester si les participants seront capables d'identifier le propriétaire le plus probable d'une séquence donnée de battements cardiaques à partir des seuls enregistrements audio de la voix d'un locuteur .
Bien qu'exploratoire, car à notre connaissance il s'agit de la première étude qui interroge la possibilité d'extraire de l'information au sujet du RC de notre interlocuteur dans sa voix, nous posons différentes hypothèses secondaires visant à interroger les facteurs qui pourraient être impliqués dans cette capacité. Aussi,

1. Au regard de la relation causale positive entre le pitch et le RC montrée dans notre expérience tilt-test, le pitch du locuteur pourrait être un indice utilisé par les participants pendant l'expérience. Nous faisons donc varier la différence entre les pitch des locuteurs présentés dans le même essai (delta pitch), et faisons l'hypothèse que plus la différence entre les pitch des locuteurs présentés dans le même essai est importante (delta pitch), plus la précision devrait être élevée pour cet essai.
2. Il est possible qu'au lieu de sélectionner le propriétaire le plus probable des battements cardiaques représentés en utilisant des indices auditifs transitoires dans les extraits vocaux, les participants utilisent des croyances préalables sur les rythmes cardiaques probables des personnes (par exemple, sur la base de

la perception de l'âge, du sexe ou de la santé). Nous voulons tester cela en comparant les performances entre 2 conditions : l'une associant au sein de chaque paire des voix de locuteurs différents, possiblement de sexe différent (condition inter-locuteur) ; l'autre associant au sein de chaque paires de extraits de voix provenant d'un seul locuteur (condition intra-locuteur).

3. Si le pitch est un indice qui sous-tend l'exécution de notre tâche, la perturber devrait affecter le choix des participants. Cette hypothèse sera évaluée à l'aide d'une version de la tâche réalisée avec pour stimuli des enregistrements de voix dont le pitch a été artificiellement manipulé.
4. Enfin, de la même manière que pour le pitch, nous observons l'effet de la différence entre les intervalles de battements cardiaques des locuteurs présentés (delta RC). Plus la différence entre les rythmes cardiaques des locuteurs est grande, plus la précision devrait être élevée pour cet essai.

5.2.2 Matériel et méthodes

5.2.2.1 Procédure

Cette expérience se compose de trois conditions : les conditions *inter* et *intra* qui sont deux versions du même paradigme avec des stimuli différents, et la condition *filtre vocal*, conçue avec des extraits de voix dont le pitch a été manipulé artificiellement. Les participants ont tous pris part aux trois conditions dont l'ordre était contrebalancé en intra-participant. Le plan expérimental était donc le suivant : un groupe de participants, trois conditions (condition inter locuteurs, condition intra locuteurs, condition stimuli filtrés ou artificiels), en mesures répétées.

1. **Condition inter locuteur**

   Lors de cette expérience nous présentons à chaque essai une paire de vidéos de 5 secondes composées chacune de l'extrait de voix d'un locuteur associé à un carré lumineux changeant de couleur du rouge au noir en suivant le rythme cardiaque correspondant à l'extrait vocal de l'un des deux locuteurs (voir photo de la tache en Fig. 5.4.

   Parmi les vidéos d'une même paire, seul l'extrait de voix diffère, le RC est fixe (celui d'un des deux extraits choisi au hasard).

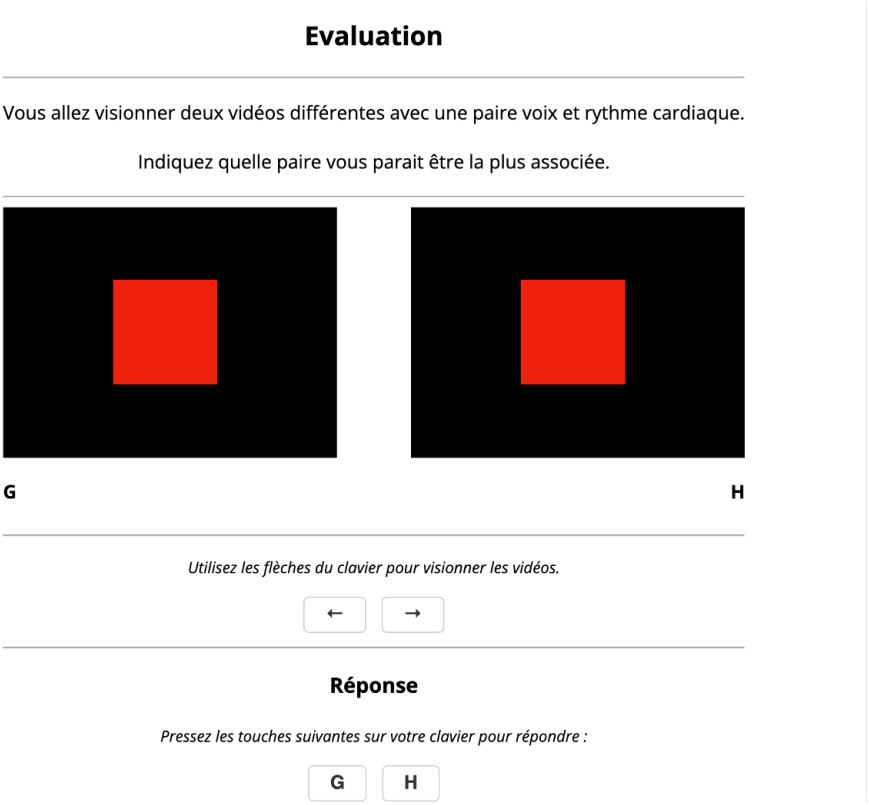


FIGURE 5.4 – Illustration de notre adaptation de la tâche "The Other's Heart Rate". A droite : extrait vocal du locuteur A diffusé en même temps que le RC correspondant. A gauche : extrait vocal du locuteur B diffusé en même temps que le RC correspondant à l'extrait du locuteur A. Le participant doit choisir la meilleure association voix-RC entre celle de droite et celle de gauche.

Tous les participants ont pour consigne pour chaque essai de choisir *« la meilleure association voix-RC »*. Après la présentation des deux vidéos, les participants doivent alors entrer leur choix, sans limite de temps et avec la possibilité de relancer les stimuli autant de fois que nécessaire. Une fois la réponse saisie, les participants sont invités à indiquer leur degré de confiance dans celle-ci sur une échelle allant de 0 (pas du tout confiant) à 10 (absolument confiant). L'essai suivant présente une nouvelle paire de stimuli. Avant le début des expériences, tous les participants effectuent deux essais d'entraînement. Après cela, 90 essais leur sont présentés, avec une petite pause de 5mn au milieu des essais. L'ordre des 90 essais est randomisé pour chaque participant. La tâche dure environ 35 minutes.

2. **Condition intra locuteur**

   Afin de nous assurer que le jugement de RC dans la voix ne repose pas sur des informations relatives à l'identité du locuteur comme son âge, son sexe ou sa corpulence, nous avons conçu une version de la tâche pour laquelle les extraits vocaux d'un même essai sont choisis chez le même locuteur. Mis à part ce changement concernant l'identité du locuteur, la tâche reste la même qu'en condition *inter-locuteur*. Les participants se voient toujours présenter deux vidéos de 5 secondes composées d'un extrait vocal et d'un carré imitant la pulsation cardiaque correspondant à l'un des deux extraits vocaux. Entre les 2 vidéos de la même paire, le seul élément qui diffère est toujours l'extrait vocal, le RC reste le même (celui d'un des 2 extraits de la paire, pris au hasard). Le reste de l'expérience notamment la consigne donnée au participant est strictement identique à la condition inter décrite ci dessus.

3. **Condition Filtre Vocal**

   Enfin, compte tenu de nos hypothèses sur le pitch, nous avons souhaité tester dans quelle mesure ce paramètre spécifique pouvait rendre compte de la détection de RC. A partir des mêmes extraits qu'utilisés en condition *intra-locuteur*, nous avons donc conçu une troisième condition impliquant des stimuli dont le pitch avait été modifié (voir section 5.2.2.2). Dans cette version, la tâche est différente de celle demandée dans les deux conditions précédentes : les participants se voient toujours présenter deux vidéos de 5 secondes, mais cette fois-ci elles sont toutes deux composées du même extrait vocal dont le pitch a été modifié, et c'est la pulsation cardiaque associée à l'enregistrement vocal qui varie. Dans un cas, il s'agit de la pulsation correspondant à l'enregistrement naturel à partir duquel avait été crée notre nouveau stimulus transformé ; dans l'autre, la pulsation cardiaque associée à l'enregistrement correspond à la nouvelle valeur de pitch de l'enregistrement présenté. La consigne reste la même : tous les participants ont pour instruction d'écouter ces paires d'audio et d'indiquer à qui appartient la pulsation cardiaque en choisissant la meilleure combinaison "extrait vocal-pulsation cardiaque". Après avoir écouté les deux audios au moins une fois, les participants indiquent alors leur réponse et le degré de confiance associé de la même manière qu'expliqué pour les deux précédentes conditions.

L'ensemble de l'expérience a été codé avec la librairie `jsPsych` 7.3.0 (©de Leeuw, 2015-2023), et hébergé sur une page `GitHub` (©Microsoft Corporation, 2008-2023). Le format de la vidéo est MPEG-4 AAC, H.264, 320 × 240 et un canal pour l'audio. Dans les trois conditions, l'ordre des stimuli est randomisé entre les participants (avec l'option randomise trials de `jspsych`). L'ordre des conditions (intra locuteurs, inter locuteurs et filtre) était quant à lui contrebalancé entre les participants.

5.2.2.2 Sélection et création des stimuli :

Création des stimuli :

Chaque stimulus est constitué de l'association d'un extrait de voix et d'un carré lumineux clignotant représentant un RC. Afin de créer ces stimuli, nous utilisons les enregistrements audio et cardiaques issus de l'expérience tilt-test (Section 5.1). Pour les besoins de la présente étude, nous avons coupé les enregistrements de production vocale pour ne conserver que les cinq premières secondes des vocalisations ainsi que les 100ms les précédant, créant de nouveaux fichiers sons de 5.1 secondes. Ce choix est motivé par le fait d'une part de créer des stimuli homogènes (même durée de vocalisation) et d'autre part qui ne soient pas trop longs à écouter. Les enregistrements PPG ont également été coupés en conséquence. Ces deux signaux physiologiques ont été analysés afin d'extraire pour chaque période de 5 secondes de vocalisation :

— le pitch moyen du signal vocal

— la FC moyenne du signal PPG.

A partir de là, nous avons ensuite crée des stimuli vidéos constitués de ces extraits vocaux auxquels s'ajoute une visualisation du rythme cardiaque. Les stimuli vidéos ont été créés avec le logiciel `ffmpeg` (© Fabrice Bellard, FFmpeg team, 2000-2023) à partir des fichiers audios de 5.1 secondes. La vidéo est encodée en 25 images par secondes et est constituée d'un fond noir noir où apparaît régulièrement un rectangle central rouge pendant 320ms (8 frames) symbolisant l'apparition d'un pic systolique.

Sélection des extraits vocaux :

Parmi les enregistrements vocaux et cardiaques dont nous disposions, seuls certains ont été sélectionné de manière à constituer les 90 essais de chacune des 3 conditions, selon la procédure suivante.

Tout d'abord, et indépendamment de la condition expérimentale, afin de ne considérer que des extraits représentatifs de la distribution statistique de paramètres

acoustiques et physiologiques enregistrée dans l'expérience « tilt-test », et exclure de possibles erreurs d'enregistrements ou « outliers » statistiques, nous avons restreint notre corpus d'enregistrements aux extraits ayant des durées comprises entre le 20e et le 80e percentile (6.0 - 9.0 sec), un pitch moyen entre le 20e et le 80e percentile (120.0 - 260.0 Hz) et un bpm moyen entre le 10e et 90e percentile (60.0 - 150.0 bpm).

Ensuite, pour sélectionner les paires d'enregistrements vocaux appariés dans la condition « inter locuteur », nous avons considéré toutes les paires possibles d'enregistrements de locuteurs distincts parmi ce corpus réduit (10311 paires possibles), et sélectionné seulement parmi celles-ci les paires d'enregistrements dont la différence entre leurs deux pitchs d'une part, et leurs deux rythmes cardiaques d'autre part, excédaient une valeur représentative du seuil de détectabilité humaine. Pour le pitch, la littérature psychophysique (Bachem, 1937) indique une JND (just-noticeable difference) sur des sons complexes comme la voix égale à 1-2% de la valeur en Hz, soit 2Hz pour des sons à 100Hz, et 4Hz pour des sons à 200Hz (soit environ 40% de demi-ton,ou 40 *cents*). Pour le RC, l'étude de Galvez-Pol Galvez-Pol et al. (2022) suggère des intervalles inter-battements supérieurs à 60ms, soit 4bpm à 60bpm et 13bpm à 120bpm. D'autre part, une étude sur la perception de différence de période dans des signaux de communication lumineux ((Laxar et al., 1990) mesure une JND de 0.25Hz, soit 15bpm. Nous avons donc retenu les paires associant des enregistrements dont le pitch diffère d'au moins 40 cents, et dont le RC différe d'au moins 10 bpm. Ce premier filtrage réduit le nombre de paires possibles à 7035.

Pour sélectionner les 90 paires finales, et afin de contrôler la difficulté perceptive de l'expérience, nous avons ensuite séparé ces 7035 paires possibles selon 3 catégories d'écart de pitch (de 0 au 33e percentile, soit 40 - 342 cents : *low-delta-pitch* ; du 33e au 66e percentile, soit 342 - 889 cents : *medium-delta-pitch* ; et du 66e au 100e percentile, soit 889-1331 cents : *high-delta-pitch*), et 3 catégories d'écart de RC (de 0 au 33e percentile, soit 10-20bpm : *low-delta-bpm* ; du 33e au 66e percentile, soit 20-35bpm : *medium-delta-bpm* ; et du 66e au 100e percentile, soit 35-88bpm : *high-delta-bpm*). Cette organisation donne 9 cellules de combinaisons possibles (*high-delta-pitch* × *high-delta-bpm* : 881 paires ; *high-pitch* × *mid-bpm* : 900p. ; *high-pitch* × *low-bpm* : 610p. ; *mid-pitch* × *high-bpm* : 1009p. ; *mid-pitch* × *mid-bpm :* 687p. ; *mid-pitch* × *low-bpm* : 624p. ; *low-pitch* × *high-bpm* : 501p. ; *low-pitch* × *mid-bpm* : 733p. ; *low-pitch* ×

low-bpm : 1086p.), dans lesquelles nous avons résolu de sélectionner 10 paires par cellule.

Enfin, afin de faire cette sélection, et pour éviter un éventuel effet « confondant » de paires de locuteurs qui n'apparaîtraient que dans certaines cellules (ex. toutes les paires homme-femme qui tomberaient par hasard dans les catégories *high-delta-pitch*), nous avons sélectionné les paires finales parmi celles correspondant à des paires de locuteurs qui contribuent à des stimuli dans au moins 6/9 cellules différentes. La sélection finale comprend donc 90 paires d'enregistrements, issus de 8 paires de locuteurs (Homme 5 - Homme 10 : 8 paires ; H5 - H13 : 12p. ; H5 - Femme 22 : 18p. ; H6 - H10 : 6p. ; H6 - H13 : 9p. ; H10 - H11 : 6p. ; H11 - F22 : 13p. ; H13 - F22 : 18p.) et répartis équitablement en 3 niveaux d'écart de pitch et de bpm.

Pour sélectionner les paires d'enregistrements vocaux appariés dans le bloc « intra-locuteur », nous avons utilisé la même logique. Le corpus fournit un total de 567 paires possibles intra-locuteur, réduit à seulement 94 paires possibles après filtrage des valeurs inférieures aux seuils de perception de pitch et de RC (réduit pour ce bloc à 30 cents et 9 bpm). Parmi ces 94 paires, nous avons sélectionné 90 paires de manière à les répartir le plus équitablement possible entre 3 catégories d'écart de pitch (*low-delta-pitch* : 30 - 56 cents ; *medium-delta-pitch* : 56 - 105 cents ; *high-delta-pitch* : 105-291 cents) et 3 catégories d'écart de bpm (*low-delta-bpm* : 9 - 12 bpm ; *mid-delta-bpm* : 12 - 18 bpm ; *high-delta-bpm* : 18 - 56 bpm) - noter que les écart de pitch et de RC entre enregistrements du même locuteur sont nécessairement moins variables qu'entre enregistrements de locuteurs différents. Les 90 paires d'enregistrements finales sont issues de 11 locuteurs différents (6 hommes et 5 femmes - H2 : 5 p. ; H3 : 3p. ; H5 : 12p. ; H10 : 2p. ; H12 : 2p. ; H13 : 7p. ; F14 : 15p. ; F15 : 2p. ; F 17 : 7p. ; F12 : 2p. ; F22 : 17p.).

Enfin, pour sélectionner les paires d'enregistrements de RC appariés dans la condition « filtre », nous avons repris les 90 paires du bloc "intra-locuteur". Pour chacune des paires, composée de deux enregistrements vocaux du même locuteur, nous avons sélectionné au hasard un des 2 enregistrements, calculé son pitch, et manipulé l'autre enregistrement avec une transformation de pitch (pitch shifting) de façon à lui attribuer le pitch du premier enregistrement. Par exemple, si une paire intra-locuteur est composée de 2 enregistrements de pitch à 120Hz et 130Hz, nous avons sélectionné au hasard l'enregistrement à 120Hz, et avons manipulé sa hauteur de

façon à ce qu'elle atteigne 130Hz. Les 90 paires transformées finales sont composées de ce seul enregistrement transformé (i.e. enregistrement 1 modifié au pitch de l'enregistrement 2, ou bien enregistrement 2 modifié au pitch de l'enregistrement 1), et des 2 signaux cardiaques d'origine (correspondant à l'enregistrement 1 et à l'enregistrement 2). La logique expérimentale de cette 3e condition est que si le pitch de l'enregistrement est un indice utilisé par l'auditeur pour attribuer un RC à une voix, alors élever le pitch de l'enregistrement 1 pour le faire ressembler à celui de l'enregistrement 2 devrait conduire l'auditeur à lui apparier le signal cardiaque de l'enregistrement 2, plutôt que celui de l'enregistrement 1.

5.2.2.3 Participants

N= 29 participants (15 femmes), tous résidents français, ont été recrutés au Centre Multidisciplinaire des Sciences Comportementales INSEAD/Sorbonne Université parmi une population composée principalement d'étudiants universitaires.

Tous les participants répondaient aux critères suivants :

— Critères d'inclusion : français comme langue maternelle.

— Critères d'exclusion : déficience visuelle ou auditive (une déficience visuelle corrigée par des lunettes ou des lentilles de contact est acceptée), trouble psychiatrique (y compris dépression, anxiété, autisme même avec des médicaments en cours), maladie cardiovasculaire, pathologie de la voix (bégaiement, etc.), traitement pharmacologique altérant la fonction cardiaque (ex. bêta-bloquant) même si aucune mesure ne sera effectuée.

5.2.2.4 Ethique

L'expérience qui a reçu l'autorisation du comité d'éthique de l'INSEAD/Sorbonne Université à la date du 21 juillet 2022, s'est tenue au Centre Multidisciplinaire des Sciences Comportementales INSEAD/Sorbonne Université en Mai 2023.

5.2.2.5 Justification de la taille de l'échantillon :

Sur la base de la méthodologie de Galvez-Pol et al. (2022), nous avons effectué une analyse de puissance a priori à l'aide de `G*Power` (version 3.1 ; © G*Power Team, 2007-2023) afin de déterminer la taille de l'échantillon nécessaire pour détecter

Tâche	Congruence	Nombre d'essais
Inter	0	70
	1	20
Intra	0	19
	1	71
Filtre	0	19
	1	71

TABLE 5.1 – Catégorisation des essais en congruent non congruent

une différence significative dans les performances par rapport au niveau de hasard de 0.5. Notre logique était que pour qu'un effet détecté soit significatif sur le plan comportemental/social, il devait avoir une taille d'effet importante (d de Cohen > 0.8). Nous voulions minimiser le taux d'erreur de type II et de type I, et avons donc fixé le bêta et l'alpha à 0.05. Une taille d'échantillon d'au moins 23 participants était l'exigence minimale pour une puissance de 0.95 (c'est-à-dire un bêta de 0.05), avec un alpha de 0.05. Sur la base de cette analyse, nous avons recruté 29 participants.

5.2.2.6 Analyses

Recatégorisation des essais :
Lors de notre précédente expérience (Section 5.1), nous avons mis en évidence une relation linéaire positive entre le pitch de nos locuteurs et leur rythme cardiaque. Cette relation décrit un effet moyen, qui n'est pas retrouvé au sein de tous nos extraits. Or, pour la présente expérience, le respect de cette relation n'a pas constitué un critère de choix de nos extraits. Pour certains d'entre eux, la relation entre les deux paramètres sera inverse. Par conséquent, selon que les extraits respectent ou non cette relation linéaire positive, nous les catégorisons respectivement sous les labels "congruent" et "non congruent", ceci pour chacune de nos conditions (inter, intra, filtre). Le nombre d'essais congruents et non congruents pour chaque condition est résumé dans le tableau 5.1 ci-après.
Nous constatons que les essais ne sont pas répartis de manière homogène entre les conditions congruente et non congruente dans les différentes tâches. Cela correspondrait à une variabilité que nous pouvons retrouver en conditions "naturelles" ou hors laboratoire.

Ajout d'une catégorie de bonnes réponses qui correspondant à « heuristique voix-coeur » :
Considérant la présence d'essais non congruents dans notre expérience, nous avons choisi d'ajouter une condition bonne réponse qui correspond à une réponse correcte à l'égard du modèle pitch-RC que nous avons mis en évidence précédemment. Cette réponse est appelée "predicted response" en référence à la théorie de prédiction qui nous amène à tester l'existence d'une règle sous-jacente au choix de les participants. Pour les conditions inter et intra locuteurs (deux sons présentés, un RC), les essais incongruents correspondent à ceux pour lesquels le pitch le plus bas est associée au bpm le plus haut, la réponse prédite correcte serait donc inverse par rapport à celle correspondant à l'association pitch-RC réellement observée. Pour la condition « filtre » (un son présenté et deux RC), les essais incongruents correspondent à ceux pour lesquels le RC le plus bas est associé au pitch augmenté (par le filtre). La réponse prédite correcte serait donc inverse par rapport à la situation observée en réalité.
En résumé, la réponse prédite est la même que la réponse correcte pour les essais congruents, et inversée pour ceux incongruents.

Variables manipulées :
Les variables dont nous testons les effets sont les suivantes :

- la tâche : condition intra-locuteur (extraits du même locuteur), condition inter-locuteur (extraits de deux locuteurs différents) , condition filtre (extraits filtrés provenant du même locuteur transformé avec filtre pitch)
- delta bpm : la différence de RC entre les deux extraits de locuteurs présentés dans un essai (répartis en trois niveaux : high, mid, low)
- delta pitch : la différence de pitch entre les deux extraits de voix présentés dans un essai (répartis en trois niveaux : high, mid, low)

Analyses statistiques :
Les analyses statistiques que nous présentons sont faites en `Python` (© Python Software Foundation, 1991-2023) avec les modules `seaborn` (© M. Waskom, 2021-2023) `pingouin` (© R. Vallat, 2018-2023) et `pymer` (© E. Jolly, 2017-2023). Dans toutes les conditions, nous calculons le pourcentage de réponse correcte en moyennant les hits de chaque essai ainsi qu'une mesure de sensibilité (d') pour chaque condition et type de réponses correctes. Nous comparons ensuite ces pourcentages et mesures

de sensibilité au niveau de la chance (0.5) en appliquant un test T de Student pour lequel nous donnons la valeur du t et sa p-value.
Nous testons également l'effet des conditions expérimentales inter-locuteurs, intra-locuteurs et « filtre » sur les pourcentages de réponses correctes et les d' en utilisant des modèles linéaire mixte (LMM). Pour les analyses de régression, `pymer` utilise le module `lmer` qui estime le modèle testé à l'aide de la méthode de vraisemblance restreinte ou résiduelle (acronyme anglais REML) et les p values ainsi que les degrés de libertés (DF) sont estimés par l'approximation de Satterthwaite. Qu'il s'agisse de modèle de régression linéaire simple ou mixte, nous présentons l'équation du modèle (dans sa syntaxe *R*), et la pertinence de chaque paramètre du modèle sera illustrée par les résultats du test T de Student de nullité statistique du coefficient associé à chaque paramètre (H0 : $\beta = 0$, contre H1 : $\beta \neq 0$) sous la forme de la valeur du coefficient β d'estimation du paramètre, la valeur du t et sa p-value.

5.2.3 Résultats

5.2.3.1 Il semble difficile d'affirmer que les individus sont capables de déterminer le RC de leur interlocuteur à travers leur voix

Notre expérience compte trois conditions expérimentales inter-, intra-locuteur et filtre. Nous commençons donc nos analyses par observer les performances des participants pour chacune des conditions expérimentales.
En condition inter-locuteur, la proportion de réponses correctes se situe entre 0.37 et 0.63 et la moyenne parmi les participants est de 0.49 (SD=0.07). Ce pourcentage n'est pas statistiquement supérieur au hasard [$t(28)$ = - 0.53 ; p =.60]. La sensibilité *d'* se situe entre -0.68 et 0.68 selon les participants (M=-0.03 ; SD= 0.34). La valeur moyenne n'est pas non plus statistiquement supérieure au hasard [$t(28)$ = - 0.53 ; p = .61, 95% CI].
Dans la condition intra-locuteur la proportion de réponses correctes se situe entre 0.38 et 0.57, avec un score moyen (M=0.49 ; SD= 0.05) non statistiquement supérieur au hasard [$t(28)$ = - 0.26 ; p = .79]. Le d' moyen (M=-0.01 ; SD= 0.26) se situe entre -0.63 et 0.37 et n'est pas non plus statistiquement supérieur au hasard [$t(28)$ = - 0.22 ; p= .82, 95% CI]. En condition filtre, la proportion de réponses correctes se situe entre 0.44 et 0.62. Le score moyen (M=0.52 ; SD= 0.04) faible, mais statistiquement

supérieur au hasard [$t(28)=2.52$, $p=.017$]. De même, le d' (M=0.1 ; SD= 0.22) se situe entre -0.29 et 0.62 et est également lui-aussi supérieur au hasard [$t(28)=2.56$, $p=.016$].
Les résultats dans les trois conditions expérimentales semblent témoigner de la difficulté des participants à réaliser les tâches. Les performances en conditions inter et intra-locuteur ne permettant pas de distinguer les choix réalisés du hasard. Si les performances en condition filtre sont quant à elles meilleures qu'au hasard, la sensibilité moyenne de 0.1 semble également refléter la difficulté de l'exercice (Figure 5.5.

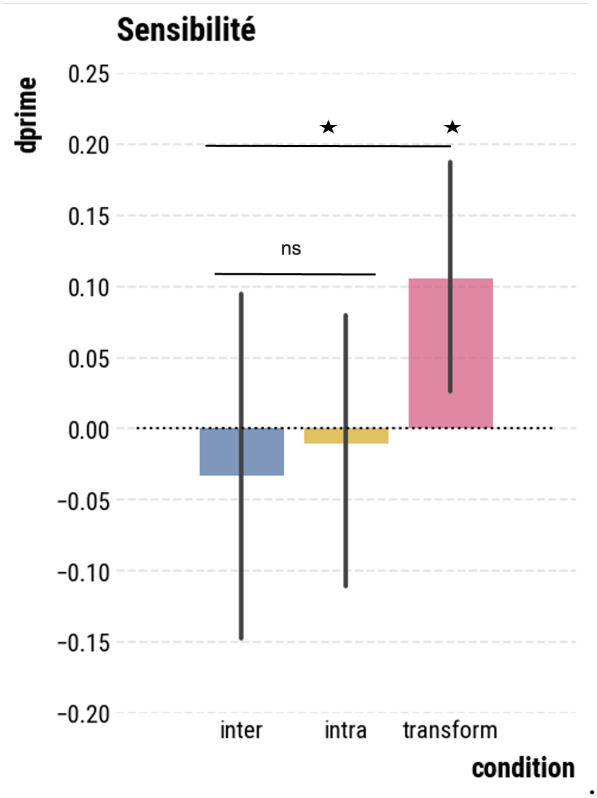


FIGURE 5.5 – Distribution des mesures de sensibilité pour chaque condition expérimentale. Les labels inter, intra, et transform sur le graphique renvoient respectivement aux conditions inter-locuteur, intra-locuteur et filtre dans le texte. Les astérisques indiquent des différences significatives et les barres d'erreurs correspondent à des intervalles de confiance de 95%.

Les deux mesures proportion de réponses correctes et sensibilité étant fortement corrélées ($r=1$, $p<.001$), nous choisissons donc de reporter uniquement les valeurs de sensibilité pour la suite des résultats afin d'en faciliter la lecture.

5.2.3.2 Le type de tâche réalisée pour tester la possibilité d'inférer le RC dans la voix a une influence sur les performances des participants

Notre expérience regroupe trois conditions (inter, intra et filtre). Dans les deux premières conditions, la tâche demandée au participant est de choisir parmi deux extraits vocaux celui qui correspond le mieux à un RC donné. Par contre, en condition filtre, nous avons évalué la possibilité d'inférer le RC dans la voix en demandant aux participants de choisir parmi deux RC celui qui correspondait le mieux à la voix présentée. Afin de savoir si la différence de tâche demandée aux participants influence le jugement, nous testons l'effet de la variable condition sur la sensibilité au moyen d'un modèle linéaire mixte :

$$\text{sensibilité} \sim \text{condition} + (1|\text{participant\_id}) \tag{5.4}$$

Les résultats de ce modèle montrent que la tâche correspondant à la condition filtre tend à être mieux réussie [$beta$= 0.139 ; t(84) = 1.898 ; p = .061], une différence qui n'est pas retrouvée entre les conditions inter et intra-locuteurs [$beta$= 0.022 ; t(84) = 0.303 ; p = .763].
La différence de tâche demandée au participant entre les conditions inter et intra d'une part et la condition filtre de l'autre semble donc avoir un effet sur les performances, favorisant la tâche qui met en jeu deux bpm différents pour un unique extrait vocal au sein du même essai.

5.2.3.3 Les participants utilisent une heuristique *pitch* $\leftrightarrow$ *bpm* pour inférer le RC dans la voix, même lorsqu'elle ne correspond pas à la réalité

Lors de notre précédente expérience tilt-test (Section 5.1), nous avons mis en évidence une relation positive entre le pitch des locuteurs et leur rythme cardiaque. Bien que tous les stimuli sélectionnés pour la présente expérience ne respectent pas forcément cette tendance moyenne, il semble plausible que les participants utilisent cette heuristique *pitch* $\leftrightarrow$ *bpm* pour inférer le RC chez autrui, même lorsqu'elle ne correspond pas à la réalité.
Pour tester cette possibilité, nous séparons les essais en essais congruents et non-congruents avec cette heuristique (voir Section 5.2.2.6), et comparons les performances des participants sur ces 2 catégories d'essais.

Pour la première tâche (choisir la voix parmi deux), en conditions inter et intra, nous observons que les d' qui étaient inférieurs à 0 pour tous les essais confondus, deviennent positifs en moyenne lorsqu'il s'agit uniquement des essais congruents. A l'inverse, pour les essais non congruents, les d' sont tous négatifs. Pour la seconde tâche également (choisir le bpm parmi 2), en condition filtre, les valeurs de d' positives sur le total des essais le restent en condition congruente mais deviennent négatives en condition non congruentes. (Figure 5.6)

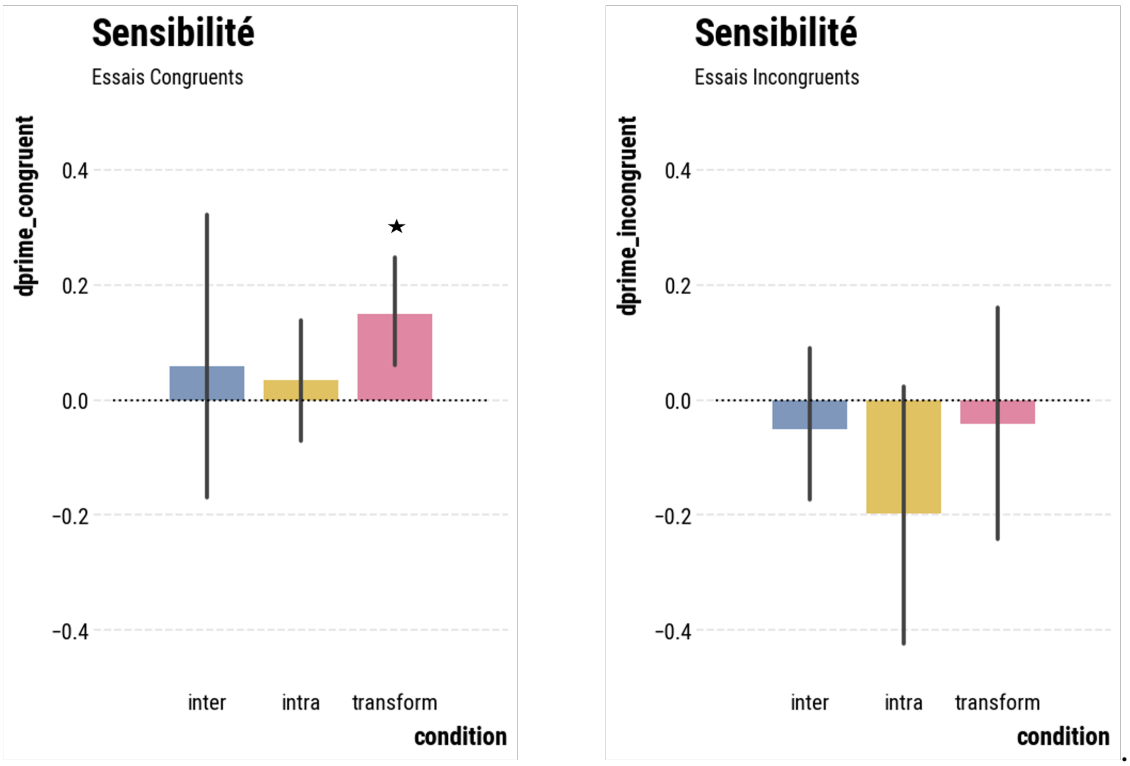


FIGURE 5.6 – Distribution des mesures de sensibilité pour chaque condition expérimentale, à gauche pour les essais congruents, à droite pour les essais incongruents au regard de l'heuristique. Les labels inter, intra, et transform sur le graphique renvoient respectivement aux conditions inter-locuteur, intra-locuteur et filtre dans le texte. Les astérisques indiquent des différences significatives et les barres d'erreurs correspondent à des intervalles de confiance de 95% autour de la moyenne.

Ces résultats, même s'ils ne sont pas tous supportés par des statistiques significatives (Table 5.2), nous invitent donc à considérer l'importance de la congruence des essais par rapport à la relation positive *pitch* $\leftrightarrow$ *bpm* mise en évidence dans notre expérience précédente. Pour évaluer à quel point les participants raisonnent conformément à

cette heuristique (càd dans le même sens que la réalité pour les essais congruents, et dans le sens inverse de la réalité, mais celui prédit par la seule heuristique *pitch* $\leftrightarrow$ *bpm*, pour les essais incongruents), nous recodons donc les valeurs prédites des essais conformément à l'heuristique (voir Section 5.2.2.6) et calculons un *d'* « heuristique » dans les 3 conditions. Comme observé dans la (Figure 5.7), les valeurs moyennes du nouveau *d'* « heuristique » sont maintenant toutes positives (inter : 0.047, sd= 0.28 ; intra= 0.062, sd= 0.27 ; filtre :0.13 , sd=0.24). Les résultats des test de Student sur ces nouvelles valeurs sont non significatives pour les condition inter ($t(28)=0.89$, $p=0.38$) et intra ($t(28)=1.24$, $p=0.22$) mais le sont pour la condition filtre ($t(28)=2.83$, $p=0.0085$).

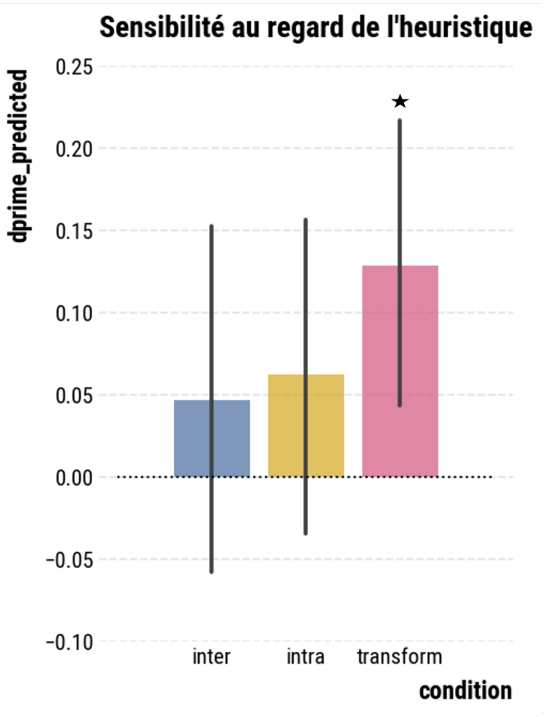


FIGURE 5.7 – Sensibilité recodée au regard de l'heuristique pour chaque condition expérimentale. Les labels inter, intra, et transform sur le graphique renvoient respectivement aux conditions inter-locuteur, intra-locuteur et filtre dans le texte. Les astérisques indiquent des différences significatives et les barres d'erreurs correspondent à des intervalles de confiance de 95% autour de la moyenne.

Il apparaît donc que les participants raisonnent conformément à une heuristique correspondant à la relation pitch b̃pm mise en évidence dans notre étude tilt-test (Section 5.1), et privilégie cette heuristique à la réalité quand elle est non-congruente.

Condition	Congruents	Non congruants
Inter	t(28)=0.46, p=0.65	t(28)=-0.75, p=0.46
Intra	t(28)=0.61, p=0.55	t(27)=-1.66, p=0.11
Filtre	**t(28)=3.16, p=0.0038** *	t(28)=-0.38, p=0.71

TABLE 5.2 – Résultats des tests de Student testant la sensibilité par rapport au hasard en tenant compte du statut congruent-non congruent des essais pour chaque condition. En gras, les statistique significatives au niveau $\alpha = 0.05$

Condition	Low delta pitch	Mid delta pitch	High delta pitch
Inter	t(28)=0.40, p=0.69	t(28)=-0.50, p=0.62	t(28)=1.45, p=0.15
Intra	t(28)=0.25, p=0.81	t(28)=-0.30, p=0.76	**t(28)=2.65, p=0.013**

TABLE 5.3 – Résultats des tests de Student testant la sensibilité recodée au regard du respect de l'heuristique par rapport au hasard en fonction du niveau de différence de pitch entre les RC présentés. En gras, les statistiques significatives.

La performance effective à la tâche dépend donc crucialement de la proportion d'essais congruents parmi les stimuli sélectionnés : les meilleurs performances à la tâche en conditions intra et filtre pourraient donc ainsi être expliquées par une plus grande proportion d'essais congruents avec l'heuristique utilisée par les participants (70/20, voir Tableau 5.1)

5.2.3.4 Influence de delta pitch et delta bpm

Enfin, ce jugement fondé sur une heuristique pitch$\tilde{b}$pm semble se faire d'autant plus facilement que les alternatives qui leur sont présentées lors des essais sont tranchées. Afin de mieux comprendre ce qui sous tend l'application de cette heuristique, nous observons l'effet des écarts de pitch et de bpm entre les stimuli présentés sur la sensibilité.

Nous observons d'une part que pour la tâche demandée en conditions inter et intra, nécessitant de distinguer entre 2 voix, une différence de pitch élevée entre les deux extraits vocaux présentés est associée à des valeurs de « predicted response » plus élevées, au point de devenir statistiquement différentes du hasard pour les essais high pitch en condition intra (t(28)=2.65, p=.013, Table 5.3)

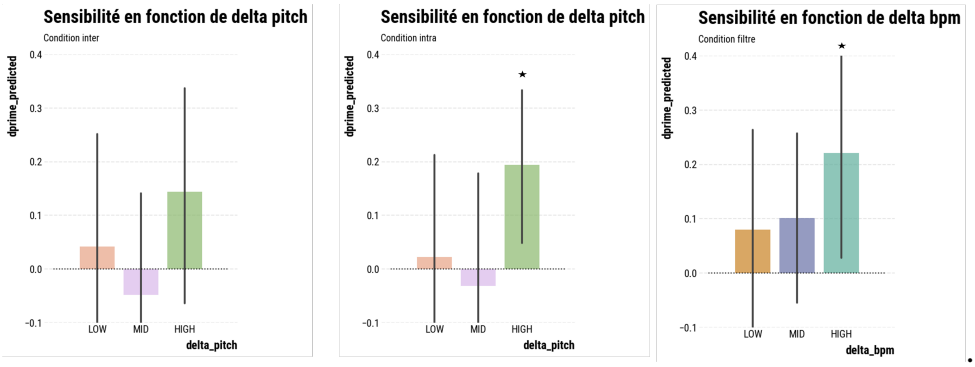


FIGURE 5.8 – Sensibilité recodée au regard de l'heuristique en fonction des niveaux de différence de pitch (LOW, MID, HIGH) entre les deux extraits présentés (en condition inter-locuteur à gauche et en condition intra-locuteur au centre) et en fonction des niveaux de différences de RC (LOW, MID, HIGH) entre les deux RC présentés (en condition filtre à droite). Les astérisques indiquent des différences significatives et les barres d'erreurs correspondent à des intervalles de confiance de 95% autour de la moyenne.

D'autre part, pour la condition filtre nécessitant de distinguer entre 2 RCs, des valeurs élevées de delta bpm sont associées à une sensibilité qui, elle aussi, se distingue des niveaux du hasard ($t(28)=2.12$, $p=0.04$, Table 5.4) comme l'illustre la figure 5.8.

Condition	Low	Mid	High
Filtre	$t(28)=0.81$, $p=0.42$	$t(28)=1.19$, $p=0.24$	**$t(28)=2.12$, $p=0.04$**

TABLE 5.4 – Résultats des tests de Student testant la sensibilité recodée au regard du respect de l'heuristique par rapport au hasard en fonction du niveau de différence de bpm entre les extraits présentés en condition filtre. En gras, les statistiques significatives.

5.2.4 Discussion

La seconde étude de ce chapitre explore la possibilité que le rythme cardiaque, signal physiologique témoin de l'état psychophysiologique d'une personne (Beauchaine, 2015 ; Thayer et al., 2012), puisse être identifié dans la production vocale des individus, de manière similaire à ce qui a été démontré dans le cas des visages (Galvez-Pol et al., 2022). Pour ce faire, nous avons conçu une nouvelle tâche expérimentale, pensée comme une adaptation du protocole de Galvez-Pol et al. (2022) au cadre de l'audition de stimuli vocaux. Nous avons donc testé la possibilité d'identifier correctement à quel extrait de voix (parmi une paire) correspond un

rythme cardiaque enregistré en cours de vocalisation. Les résultats que nous observons à cette tâche, testée lorsque les voix présentées appartiennent à deux locuteurs différents d'une part et à un seul et même locuteur enregistré à deux moments distincts de l'autre, ne permettent pas de répliquer les résultats obtenus avec des visages.

Pour autant, à notre sens, il ne semble pas que nous puissions exclure complètement la possibilité que les individus soient capables d'extraire de la voix de leur interlocuteur des informations au sujet de son rythme cardiaque. En effet, lorsque la tâche proposée consiste cette fois à choisir parmi deux RC celui correspondant le mieux à un extrait vocal donné, les participants performent mieux qu'au hasard. Ce résultat nous amène à considérer la voix comme véhicule potentiel d'informations au sujet du signal cardiaque que les individus seraient à même d'extraire et d'utiliser dans certaines conditions.

Il convient alors de préciser que cette seconde tâche (qui propose de choisir entre deux bpm) comporte la spécificité de présenter des extraits vocaux dont le pitch a été manipulé. Cette manipulation vise à questionner l'utilisation de l'heuristique liant pitch et RC (augmentation de bpm associée à augmentation du pitch) mise en évidence dans notre étude de tilt-test (Section 5.1) pour inférer le RC d'un individu à partir de sa voix. Rappelons alors que de manière cohérente avec notre question, nous considérons pour cette tâche un jugement comme correct lorsque le bpm choisi est celui correspondant au pitch modifié. Les résultats en condition « filtre » semblent indiquer que les participants utilisent bien l'heuristique *pitch* $\leftrightarrow$ *bpm* pour inférer le RC à partir d'un individu à partir de sa voix.

De plus, cette heuristique semble également influencer les jugements des participants lors de la tâche adaptée du protocole de Galvez-Pol et al. (2022)(juger à quelle voix une pulsation cardiaque donnée correspond le mieux) compte tenu de l'observation qualitative de l'incidence du caractère congruent ou non des essais relativement à l'heuristique *pitch* $\leftrightarrow$ *bpm* sur les scores de sensibilité ; positifs pour les essais congruents (et lorsque la tâche est recodée au regard de l'heuristique *pitch* $\leftrightarrow$ *bpm*) et négative pour les essais non congruents. De surcroît, ces différentes évidences nous invitent à penser que même dans les situations « naturelles » pour lesquelles la

relation *pitch* ↔ *bpm* ne s'applique pas (essais non congruents), c'est elle qui prévaut pour inférer des informations cardiaques dans la voix, et conduirait à des erreurs de jugement au regard des informations réelles. Ces résultats s'inscrivent alors manifestement dans le cadre de la théorie du codage prédictif qui implique que nos modèles du monde fondés sur des priors (ici *pitch* ↔ *bpm*) étayent nos inférences au sujet des causes cachées des stimuli et peuvent dans certains cas être associés à des erreurs de prédiction.

Par conséquent et de manière attendue, il semblerait que des conditions favorables à l'utilisation des indices contenus dans notre modèle *pitch* ↔ *bpm* améliorent les jugements des participants. En effet, cette amélioration s'observe quand la différence de bpm entre les deux RC d'un même essai est la plus importante (essais high-delta-bpm en condition filtre, entre 18 et 56 bpm d'écart) et lorsque les écarts de pitch entre deux extraits vocaux relèvent de la catégorie high-delta-pitch (entre 889 et 1331 cents en inter et entre 105 et 291 cents en intra), au point d'observer en condition intra-locuteur une sensibilité significativement différente du hasard pour attribuer une voix correctement au regard de l'heuristique parmi deux extraits de voix d'un même locuteur à un RC donné.

Enfin, les résultats de notre étude indiquent que le pitch semble prévaloir sur les autres marqueurs acoustiques présents dans la voix (qui eux, ne sont pas modifiés par le filtre) pour faire des inférences sur le RC d'un locuteur. Ainsi, l'usage d'un filtre permettant de manipuler spécifiquement le pitch (pour l'augmenter ou le baisser) dans un extrait vocal conduit à ce que les participants choisissent d'associer à ce dernier le rythme cardiaque correspondant au pitch crée par le filtre indépendamment des autres paramètres de la voix. En cela, l'incidence de la manipulation de pitch dans les extraits vocaux nous semble se rapprocher de ce que nous avons pu observer concernant la perception de l'état du patient par les psychiatres, cette dernière suivant les valeurs de pitch malgré les autres indices possiblement présents dans la voix des patients.

Le fait qu'une heuristique « vraie » puisse finalement induire en erreur les individus, en plus de renvoyer aux résultats constatés chez les soignants dans notre expérience au sujet de la perception de l'état de gravité d'un patient 4.3.4 a notamment été montré dans le cadre de la communication médiée par IA au sujet du langage généré par IA. Ainsi, le travail de Jakesch et al. (2023) questionne la possibilité par des humains

de d'identifier des écrits comme ayant été générés par une IA ou par un autre être humain. A cette fin, les chercheurs font évaluer de courts textes de présentation de soi et montrent que les jugements humains du langage généré par l'IA sont entravés par des heuristiques telles que l'association de pronoms de la première personne, l'utilisation de contractions ou l'abord de sujets familiaux avec le langage écrit par l'homme. Ce travail démontre expérimentalement que ces heuristiques rendent le jugement du langage généré par l'IA prévisible et manipulable, ce qui permettrait aux systèmes d'IA de produire des textes perçus comme *« plus humains qu'humains »*.

Limitations :

Compte tenue de la nouveauté de notre paradigme, nous avons choisi de réduire la complexité de la tâche et de commencer par adresser notre question en utilisant des productions vocales simples (voyelles soutenues) pour support de notre tâche de perception du RC au sein de la voix d'autrui. La généralisation de résultats obtenus pour la production d'une voyelle soutenue à une situation écologique d'interaction verbale (en parole continue) n'est donc pas directe. En cela, nous nous plaçons en quelques sortes au sein des mêmes limites que Galvez-Pol et al. (2022) relèvent pour leur étude, à savoir mettre en évidence la possibilité de détecter le RC d'autrui au sein de courtes vidéos dans lesquelles les acteurs - non représentatifs de la population car jeunes et majoritairement européens - sont immobiles, ne montrent aucune émotion particulière, et ne sont engagés dans aucune action. De tels résultats ne sont donc pas nécessairement généralisables tels quels à des conditions hors-laboratoire. Dès lors, il s'agira de mener des investigations supplémentaires en condition plus proche de celle de l'interaction sociale. A cette fin, nous avons recueilli lors de l'expérience tilt-test un corpus de données parallèles collectées pendant que nos locuteurs prononçaient une phrase standardisée. L'analyse future de ces données permettra de juger de la réplication de nos résultats dans des conditions plus « naturelles ».

D'autre part, au regard de l'objet de notre thèse qui est le filtre de transformation de pitch (et de l'importance de ce paramètre acoustique dans les interactions sociales), nous nous sommes focalisés sur le pitch comme indice à partir duquel il serait possible d'extraire des informations au sujet du RC du locuteur. Même si ce paramètre de pitch semble, dans des conditions favorables (tâche adaptée, gros écart de pitch entre les stimuli) suffire à expliquer les réponses des participants, il reste probable que d'autres composantes acoustiques co-varient également avec le RC. Des analyses acoustiques

complémentaires, à partir du corpus 5.1 gagneraient à explorer ces indices acoustiques additionnels, possiblement à partir d'études systématiques de type reverse-corrélation (Pruvost-Robieux et al., 2022). Un espace possible d'investigation pourrait ainsi être celui des spectres de modulations spectro-temporelles (*rate* et *scale*) (Thoret et al., 2022), et notamment des modulations temporelles de la même échelle que celle du bpm (1-3Hz), ce qui rejoindrait certaines données déjà publiées dans la littérature sur les liens entre bpm et jitter (Orlikoff and Baken, 1989). De surcroît, la limitation de nos analyses à l'influence du seul pitch dans ce processus d'inférence du RC à partir de la voix ne nous permet pas de traiter la question d'une possible incongruence dans les informations véhiculées par d'une part le pitch manipulé et les paramètres vocaux non visés par le filtre de l'autre. Si incongruence il y avait effectivement, nous pouvons nous demander quelles seraient alors les conséquences en terme de charge cognitive dans le traitement de ces informations. La question de la charge cognitive associée aux interactions médiées par la technologie est adressée par les travaux récents de Schwartz et al. (2022) qui rapportent que les interactions via une plateforme de visioconférence (sans usage de filtre) atténuent le phénomène normal de synchronisation inter-cérébrale entre deux individus. Ainsi, en mesurant la dynamique inter-cérébrale de dyades mère-enfant en interaction en face à face ou en vidéo-chat (à la Zoom), les chercheurs observent que l'interaction en face à face suscite une plus grande synchronisation inter-cérébrale que l'interaction par vidéo-chat. Leurs résultats indiquent que la co-présence humaine est sous-tendue par des processus neurobiologiques spécifiques qui méritent d'être étudiés en profondeur. Parmi les conséquences possibles, les chercheurs suggèrent que l'effet de « Zoom fatigue » associé aux pratiques de visio-conférence puisse être relié à une surcharge des connexions inter-cérébrales plus limitées en condition Zoom. A l'instar du travail du travail de (Schwartz et al., 2022), des études examinant la spécificité possible du traitement cérébral des stimuli filtrés contribueraient également à éclairer la compréhension des effets du filtre vocal sur les mécanismes cognitifs.

5.3 Conclusion générale aux deux études

Considérant les divers résultats que nous reportons ici, il semble difficile de conclure avec certitude au sujet de la possibilité que les individus perçoivent aisément

le RC dans la voix de leur interlocuteur. Cependant, il semblerait qu'une heuristique *pitch* $\leftrightarrow$ *bpm* fondée sur des priors correspondant à ceux mis en évidence dans l'expérience tilt (Section 5.1) soit utilisée par les individus dès lors qu'ils ont à extraire dans la voix de leur interlocuteur des informations relatives à son RC (Section 5.2).

Parallèlement à la contribution biologique de cette étude, celle-ci participe de la caractérisation des effets possibles de l'utilisation d'un filtre vocal sur notre cognition sociale fine. Parmi les apports singuliers de ce travail, réside à notre avis le fait qu'il renseigne sur l'effet que le filtre vocal peut avoir sur la perception d'un signal physiologique autre que celui directement visé par la transformation : ici le RC. En outre, sachant la signification émotionnelle qui peut être attribuée au signal cardiaque, il est possible que ces inférences au sujet du RC soient intégrées au sein d'un processus d'inférence plus large visant à attribuer des émotions à un interlocuteur, celui-ci se construisant à l'aune d'une multiplicité d'indices.

L'influence du filtre de pitch nous parait d'autant plus importante à considérer qu'elle repose sur une heuristique *pitch* $\leftrightarrow$ *bpm* correspondant à la tendance statistique retrouvée dans notre expérience tilt-test et qui semble, dans certaines conditions primer sur la réalité et entraîner certaines erreurs de jugements.

En conclusion, en plus d'impacter les inférences que les soignants se font de l'état d'un patient (Section 4.3.4), le filtre de pitch agirait sur celles que l'on se fait au sujet du RC d'un individu. Le fait qu'une manipulation de pitch puisse guider certaines de nos inférences semble alors être un argument au sujet du potentiel anthropotechnique de l'objet filtre vocal que nous avions envisagé en introduction de ce manuscrit.

Déclaration de contribution

Nadia Guerouaou : Conception de l'étude, Collecte des données, , Analyse des données - statistiques, Rédaction - préparation du projet original.
Paul Maublanc (ingénieur de recherche) : Analyse des données - acoustique et cardiaque, Conception des stimuli, Analyse des données - statistiques
Matthieu Fraticelli : Design de l'expérience en ligne, Collecte des données
JJ Aucouturier (co-encadrant) : Conception de l'étude, Analyse des données - acoustique, Analyse des données - statistiques, Rédaction - préparation du projet original.
Guillaume Vaiva (co-encadrant) : Conception de l'étude, Rédaction -révision du projet.

6. Conclusion

6.1 Résumé des résultats

Cette thèse se propose de réfléchir au devenir de notre capacité de perception des émotions, façonnée par la culture, dans un monde où les objet techniques permettent dorénavant de créer et contrôler artificiellement des indices expressifs faciaux et vocaux jusqu'alors considérés comme « naturels » au moyen d'objets que nous avons appelés *filtres* dans le corps de cette thèse.

Le contexte au sein duquel ce travail de thèse a été réalisé est alors, d'une part, celui de l'émergence de technologies capables de contrôler et mimer artificiellement des signes expressifs extérieurs que nous associons généralement (en occident du moins) à des phénomènes émotionnels et, d'autre part, l'opportunité clinique d'appliquer ces outils pour la prise en charge du trouble de stress post-traumatique (TSPT). Partant de cet état de fait, nous avons proposé de réfléchir au devenir de nos interactions et en particulier de notre cognition sociale fine en lien avec l'utilisation d'un outil de filtre vocal.

Nous avons ainsi commencé notre travail en montrant au moyen d'une étude d'éthique expérimentale (Chapitre 2) que l'utilisation de filtres vocaux permettant de modifier artificiellement la tonalité émotionnelle de la voix bénéficiait d'une grande acceptabilité morale au sein d'une population de jeunes français, et que cette acceptabilité était encore meilleure dans certains contextes tels qu'un usage à visée thérapeutique ou pour atténuer des expressions liées à des émotions de valence négative.

Deux études menées au cours des séances successives de thérapie d'exposition en imagination (Chapitre 3) nous ont permis d'adresser la question de l'effet du TSPT sur la voix de patients pendant qu'ils font le récit de l'expérience traumatique qu'ils

ont vécue.
De façon plus précise, le caractère longitudinal de ces deux études nous a permis de suivre les comportements de quatre paramètres acoustiques au long de la thérapie et d'éclairer les facteurs pouvant expliquer ces comportements. Nos résultats ont mis en évidence la pertinence du *pitch*, dont la diminution au fur et à mesure de la thérapie semble particulièrement liée au processus de guérison d'une part et à la baisse de l'intensité émotionnelle éprouvée lors de la narration au fil des séances successives (reflétée par la relation entre pitch et cluster B) d'autre part. De surcroît, les données de notre deuxième cohorte TraumacoustiK extension physiologique indiquent que cette baisse du pitch avec la rémission de la symptomatologie semble être également associée au rythme cardiaque du patient. Bien que l'association entre le pitch et le RC alors observée semble s'inscrire dans un ensemble de données corrélationnelles liant ces deux variables dans la littérature chez l'individu non malade (Schuller et al., 2014), le fait qu'au sein de nos données l'augmentation du pitch soit associée à une baisse du RC n'est pas retrouvée dans la littérature (Usman et al., 2021).

Les résultats précédents et en particulier le fait qu'il existe bien des paramètres dans la voix qui suivent l'évolution du TSPT pendant l'exposition en imagination à l'ET ont alors conforté notre intuition que des transformations simples de la voix comme une hausse ou une baisse de pitch seraient à même, selon les contextes, de reproduire des comportements vocaux à haute signification personnelle. De fait, un filtre de pitch appliqué sur la voix d'un patient pourrait « simuler » le passage d'un statut malade ou guéri ainsi que l'intensité émotionnelle ressentie par le patient lors du récit de l'évènement traumatique. Dès lors, pour la suite de ce travail, nous avons examiné le filtre du pitch comme un cas d'étude pour une forme d'anthropologie cognitive de l'objet « filtre vocal » et tenté d'apporter des éléments de réponse à la question que nous avions introduite en début de manuscrit : quelles seraient les conséquences sur le plan cognitif de l'utilisation de technologies de transformation vocale pour les interactions sociales ? A cette fin, nous nous sommes penchés sur les effets d'un tel filtre sur nos perceptions.

Au chapitre 4, nous avons mené une étude perceptive examinant l'évaluation que les soignants font de l'état de santé psychologique de patients à partir de l'écoute de leur voix, ainsi que l'influence d'un filtre de pitch sur cette perception. Pour cela, nous avons demandé à des soignants de comparer des extraits de voix deux à deux,

correspondant au même moment du script traumatique, enregistré en début et en fin de thérapie. Selon les conditions, l'un des deux extraits présentés avait subi une transformation de pitch (baisse pour les extraits initiaux et hausse pour les extraits finaux, selon la relation mise en évidence dans nos cohortes cliniques du Chapitre 3). Nous avons alors montré que les soignants experts de la thérapie d'exposition en imagination sont capables de juger de l'état de guérison de patients à partir de leur seule voix, et qu'ils utilisent le pitch - plus spécifiquement une heuristique *pitch* $\leftrightarrow$ *gravité des symptômes* - pour faire ce jugement, jusqu'à parfois les induire en erreurs.

Suite à l'observation la relation *pitch* $\leftrightarrow$ *bpm* retrouvée au sein de la cohorte TraumacoustiK extension physiologique, nous avons voulu explorer plus en profondeur le lien entre ces deux paramètres en utilisant cette fois une manipulation causale du RC inspirée du protocole médical de *tilt test* (Chapitre 5). Celle-ci nous a permis de confirmer l'existence d'une relation entre ces deux variables, bien que retrouvée ici dans un sens positif. Ce résultat souligne, si besoin, l'intérêt de la manipulation causale dans la démarche expérimentale, par rapport à la seule description corrélationnelle de facteurs observés, mais non manipulés, dans un corpus (Casadevall and Fang, 2008). Dans une démarche similaire à celle adoptée au chapitre 4, nous avons ensuite testé l'influence de la manipulation de pitch sur la perception du RC chez des individus non malades, à la manière de travaux récents sur la base de visages par Galvez-Pol et al. (2022). Cette expérience nous a alors permis de constater que dans des conditions favorables (la tâche étant difficile), les participants utilisaient bien le pitch et plus précisément la relation *pitch* $\leftrightarrow$ *bpm* retrouvée lors de l'expérience tilt-test pour faire des inférences au sujet du RC à partir de la voix d'un locuteur.

Face à de tels résultats, ce travail permet à notre sens d'apporter un premier argument en faveur de notre hypothèse élaborée en introduction, qui propose l'influence du filtre vocal sur les processus d'inférence perceptive ainsi qu'un début de contribution à la question de savoir ce que l'objet filtre pourrait faire à nos cognitions sociales fines en situation d'interaction .

De fait, notre travail suggère que le pitch est un indice utilisé pour élaborer des heuristiques au sujet de deux *états cachés* selon la formulation bayésienne : l'état de gravité d'un patient d'une part chez les soignants experts de la thérapie d'exposition en imagination (Chapitre 4) et le rythme cardiaque d'un individu chez les participants sains (Chapitre 5). Bien qu'il nous semble pouvoir penser que les deux types d'inférence

ne sont pas également utilisées au sein des interactions, la capacité à percevoir le RC dans la voix semblant être une tâche difficile, nous avons pu mettre en évidence le fait qu'un filtre de pitch avait sur ces deux processus d'inférence (gravité de l'état du patient et rythme cardiaque d'un interlocuteur) un effet majeur.

Chez les soignants, nos résultats nous invitent à penser que l'heuristique qu'ils utiliseraient pour juger de l'état d'un patient à partir de sa voix serait apprise implicitement à l'écoute des séances d'exposition en imagination et bénéficierait d'un poids considérable au sein du modèle cognitif des experts. Cette heuristique semble guider leur jugement au point de les induire en erreur lorsqu'on manipule ici les extraits vocaux par un filtre de pitch. Ils jugeraient alors qu'un patient est malade alors qu'il ne l'est pas.

Chez les sujets sains, l'heuristique *pitch* $\leftrightarrow$ *bpm* est elle aussi possiblement apprise à partir de régularités extraites dans la « nature » comme nous le suggèrent les données de notre expérience tilt-test. Cette heuristique semble guider le jugement des individus dès lors qu'ils ont à choisir d'associer un RC à une voix. Cette inférence se faisant sur la base du pitch manipulé au dépit des autres indices possiblement portés par la voix, elle semble ici encore résister au réel et volontiers induire le participant en erreur. La signification émotionnelle pouvant être attribuée au signal cardiaque nous invite à considérer la possibilité que cette heuristique soit intégrée au sein d'un processus d'inférence plus large visant à attribuer des émotions à un interlocuteur.

A ce titre, il apparaît que les résultats que nous compilons dans ce manuscrit et leur inscription dans le cadre de la théorie du traitement prédictif puissent être utilisés à titre d'explenans au sujet de notre questionnement initial. De fait, ils témoignent, par l'exemple de la manipulation de pitch, de l'effet que l'objet filtre vocal pourrait avoir sur les inférences que nous faisons au sujet d'autrui. De surcroît, ils nous enjoignent à notre sens à penser les conséquences de ces effets potentiels sur nos interactions avec autrui.

Ainsi, dans son dernier ouvrage, *The experience machine*, Clark (2023) posait les questions suivantes au sujet des processus d'inférence perceptive : « quelle est la relation avec la réalité que vous percevez ? de quelle manière la façonnons-nous et, par extension, de quelle manière nous nous façonnons nous-mêmes, souvent sans le savoir ? ». A l'aune des résultats de notre travail, il nous semble que nous pouvons

étendre cette question au filtre vocal et à la façon dont, par son usage, nous pourrions alors façonner notre réalité et nous mêmes.

6.2 Potentiel anthropotechnique du filtre vocal

Le fait qu'une manipulation de pitch puisse guider certaines de nos inférences semble être un argument au sujet du potentiel *anthropotechnique* de cet objet envisagé en introduction de cette thèse. De fait, parce que son usage se fonde sur des heuristiques auxquelles les participants se fient manifestement, nous proposons de considérer le filtre vocal comme une « technologie de soi », pour reprendre l'expression utilisée par Allard (2009) au sujet du téléphone portable et en référence aux « techniques de soi » de Foucault (1983). Le façonnement du soi que permettrait le contrôle paramétrique de nos indices vocaux implique non seulement un façonnement de *l'image de soi* renvoyée à un interlocuteur, mais également et surtout (pour notre objet d'étude) un façonnement de *soi* dans le sens d'une « reprogrammation » implicite des modèles génératifs étayant la perception des états de notre interlocuteur en situation d'interaction émotionnelle (Pajon, 2012).
Dès lors, nous pouvons nous poser ici la question des effets associés au caractère anthropotechnique de cette technologie, et des conditions qui pourraient rendre possible leur observation.

6.2.1 Conditions de déploiement des effets du filtre vocal

Dans son article *Techniques de soi et technologie numériques*, Jacobs (2018) débute son analyse par la description suivante de notre siècle « *Le XXIe siècle est celui du « numérique intégral » (Stiegler, 2016a). Pas un domaine de la société, de ses secteurs d'activité et de la vie qui ne soit investi, occupé, converti et façonné par les technologies numériques* ». Cette « *mutation prodigieuse et unique* » (Jacobs, 2018) (jamais dans l'histoire des techniques l'introduction d'une technologie n'aura transformé la société aussi radicalement et avec une telle fulgurance) s'accompagne dorénavant de l'introduction de l'IA dans la communication interpersonnelle. Cette dernière aurait « *le potentiel de transformer une fois de plus la façon dont les gens communiquent, de bouleverser les présomptions relatives à l'agentivité et à la médiation, et d'introduire de*

nouvelles questions éthiques » (Hancock et al., 2020). La communication médiée par l'intelligence artificielle (*Artificial Intelligence-Mediated Communication*, AI-MC) est ainsi une communication interpersonnelle « *qui n'est pas simplement transmise par la technologie, mais modifiée, augmentée, voire générée par un agent informatique pour atteindre les objectifs de la communication* » (Hancock et al., 2020).

En plus de l'extension de la numérisphère qui caractérise notre contexte sociétal actuel, nous pouvons évoquer parmi les éléments contribuant à l'influence potentielle du filtre vocal, la charge nouvelle dévolue au numérique que Sadin (2018) décrit comme une « *puissance aléthéique, une instance dévouée à exposer l'alètheia, la vérité, dans le sens défini par la philosophie grecque antique entendu comme le dévoilement, la manifestation de la réalité des phénomènes au-delà de leurs apparences* ». Le numérique serait selon lui « *un organe habilité à expertiser le réel de façon plus fiable que nous-même* » et « *à nous en révéler des dimensions jusque là voilées à notre conscience* ». En cela, il prendrait la forme d'une « *techne logos, une entité artificielle douée du pouvoir de dire toujours plus précisément et sans délai, l'état supposé exact des choses* ». Il nous semble alors que le crédit accordé au logos charrié par le numérique pour statuer au sujet de l'état du monde qui nous entoure pourrait alors nous rendre particulièrement sensible aux effets du filtre vocal. L'idée de potentiel aléthéique est également retrouvée chez Heidegger (1958) dans la définition qu'il fait de la technique qui par essence consisterait à faire naître, à révéler. Ainsi par exemple, la technologie énergétique moderne a révélé que le Rhin était un dépôt d'énergie plutôt qu'un simple fleuve. A cet égard, l'objet technique aurait le potentiel, sinon de créer le réel, d'en révéler un sens nouveau auquel notre société accorderait un crédit tout particulier. Ceci viendrait alors s'ajouter au fait que nos études semblent mettre en évidence un poids important des heuristiques à partir desquelles le fonctionnement du filtre est réfléchi. Chez Heidegger l'analyse de cette essence de la technique moderne s'accompagne de l'observation du danger d'exploitation qu'elle charrierait.

6.2.2 Injonctions sociales à l'auto-personnalisation

Comme nous l'avons explicité, le filtre vocal pourrait s'observer comme un objet de façonnement de soi qui, d'après nos travaux décrits en Chapitre 2 dans le cas du filtre vocal émotionnel, serait particulièrement bien accepté par la population. Considérant

la valorisation par notre société contemporaine du contrôle de nos émotions et plus généralement de notre image, la bonne réception d'une technologie permettant de maîtriser parfaitement ses manifestations affectives comportementales n'est donc peut-être pas très surprenante (Clarke et al., 2003). Nous avons d'ailleurs observé que cette grande acceptabilité était associée à une tendance à la recherche de contrôle élevée chez les participants, à l'instar des utilisateurs de *face filters* sur `Instagram` dont une des motivations principales serait celle de présenter un soi idéal (Javornik et al., 2022). A ce sujet nous pouvons alors citer Illouz (2006) qui propose de « *traiter les émotions [...] comme Marx les marchandises* » et souligne que « [Les émotions] *sont façonnées par les rapports sociaux, qu'elles ne circulent pas librement et sans contraintes, que leur magie est une magie sociale, et qu'elles contiennent et condensent les institutions de la modernité* ». Dans la même veine, selon Pajon (2012), « *l'avenir qui se dessine est moins celui de la fabrication intégrale et répétitive des corps que celui d'une modulation fine et personnalisée adaptée aux exigences des marchés* ».

Dans le monde professionnel, la thèse du « travail émotionnel » de la sociologue A.R Hochschild décrit la codification de l'affichage des émotions dans ce milieu. A la suite de Goffman (1973) qui décrivait plus tôt le « travail sur le visage »[que l'on veut se donner] (*face work* en anglais), Hochschild et al. (2017) montre comment, au cours des dernières décennies, dans le cadre d'une économie capitaliste où prédominent les emplois de service, ce travail émotionnel a été orienté vers des fins marchandes par les entreprises. Elle illustre son propos par l'exemple d'une compagnie aérienne qui assigne à ses employées de « sourire comme si elles étaient réellement heureuses ». Un tel travail à l'endroit des émotions affichées serait alors associée à des difficultés chez ces mêmes employées parfois incapables de se libérer d'un sourire artificiel comme en témoigne cet extrait d'un entretien de la sociologue avec une hôtesse de chez Word Airways : « *Parfois je reviens totalement épuisée d'un long voyage, mais je me rends compte que je ne peux pas me détendre. Je passe mon temps à glousser, je jacasse, je passe des coups de fil à des amis. C'est comme si je n'arrivais pas à me libérer d'une espèce d'allégresse artificielle créée artificiellement pour me permettre de rester énergique pendant le voyage* » (Hochschild et al., 2017). En résumé, l'enjeu éthique de l'autonomie pourrait être considérablement engagé ici à l'instar du risque soulevé par le CCNE (2013) au sujet des techniques de *neuroamélioration* : « L'individu se croit libre de tout, mais en réalité il est sous l'effet d'une injonction à la performance. La

recherche éperdue d'une performance mue par le désir impérieux de progresser peut masquer la plus contraignante des aliénations ».

Pour autant, dans le domaine numérique, il y a toujours eu des mouvements non seulement critiques, mais également alternatifs, qui cherchent à utiliser la technologie pour s'opposer aux pouvoirs établis. Un des textes fondateurs de ces mouvements alternatifs est *Le Manifeste Cyborg* de Haraway (1984) dans lequel la chercheuse pose que nous serions déjà « *des chimères et des hybrides* ». En cela, nous pourrions la rapprocher du discours transhumaniste, mais ce serait faire défaut à la pensée de Haraway qui est très différente. C'est une « *pensée de l'hybridation, de la génération, dans laquelle le lien à la machine comme aux autres formes de vivant est une manière de se libérer des identités assignées (de genre, sociales, nationales, etc.) et de se produire soi-même de manière beaucoup plus diverse et variée* » (Andler, 2021). Une telle proposition nous semble alors à rapprocher de la description qui est faite de certains usages de *face filters* qui peuvent être vécues comme des expériences permettant d'expérimenter de nouvelles identités et qui sont alors associés à des états mentaux positifs chez les usagers.

6.2.3 Glissement de la norme : du normal à l'indésirable

Les résultats de notre étude éthique et particulièrement le fait que de manière inattendue, les transformations visant à renforcer les expressions positives (sourires) soient jugées moins acceptables que celles visant à réduire les expressions négatives (anxiété, colère) donne matière à envisager ici le phénomène de *glissement de la norme* classiquement abordé dans le cadre de la réflexion éthique au sujet des technologies dites d'augmentation (CCNE, 2013). Celui-ci renvoie au déplacement de la frontière entre ce qui est jugé « normal » et ce qui relèverait selon les cas du pathologique ou de l'indésirable, l'action technique pouvant facilement faire bouger les frontières comme le souligne Goffi (2009) au sujet des techniques d'augmentation biomédicales.

Les implications morales de l'innovation - c'est-à-dire la manière dont l'introduction de nouvelles technologies affecte les relations, les identités, les normes et les valeurs - sont parfois appelées *soft impact* (van der Burg, 2009). Bien qu'ils désignent des phénomènes qui sont souvent loin d'être « mous » dans le sens où ils seraient facilement modelables - de fait les normes changent lentement et de surcroît rarement à la suite

d'un effort délibéré pour les changer - le concept de *soft impact* s'est avéré précieux pour adresser des questions autres que celles qui sont plus étroitement associées aux aspects fonctionnels des technologies. Ces derniers types de problèmes, qui comprennent les risques liés à la sécurité et la santé notamment, sont pour des raisons de contraste, appelés *hard impacts* par le fait qu'ils présentent un risque d'occurrence quantifiable. L'idée principale qui sous-tend le concept de *soft impact* est donc que la technologie et la norme morale évoluent conjointement. De fait, selon Boenink et al. (2010) « *les développements technologiques ne seront pas seulement promus ou contestés en termes de principes moraux généralement acceptés (c'est bien ou c'est mal de faire ceci ou cela), mais pourraient également provoquer des débats remettant en question le paysage moral établi* ».

Le fait que nous observions dans notre étude d'éthique expérimentale que l'atténuation des émotions négatives soit encore mieux acceptée que l'augmentation du sourire pourrait ainsi préfigurer une situation dans laquelle, une large adoption de cette technologie expressive déplacerait la responsabilité morale associée à certaines émotions ou à certains comportements. Des expressions qu'il est actuellement acceptable de ne pas aisément contrôler (comme le trémolo dans la voix chez les étudiants lors d'examens oraux - Ben-Ze'ev (1997)) pourraient devenir contrôlables, et donc blâmables et soumises à la pression sociale (« *pourquoi n'avez-vous pas appliqué le filtre contrôle du stress ?* », Theriault et al. (2021)). Pour approfondir cette idée, il serait intéressant d'examiner des scénarios impliquant des expressions positives mais qui seraient exprimées de manière difficilement contrôlée (par exemple, utiliser une transformation pour éviter un fou rire à un moment non opportun) ou d'examiner comment les résultats actuels seraient modulés par les différences culturelles dans les normes d'affichage des émotions Matsumoto et al. (2008). La possibilité de modification des normes morales par l'utilisation généralisée du filtre vocal invite à explorer les *soft impacts* associés à cette technologie. Adresser cette question paraît d'autant plus nécessaire que le respect de cette nouvelle norme serait alors conditionné par l'usage et donc la possession de cet outil, pouvant alors résulter en la création d'une *classe sociale améliorée* « avec un risque évident de discrimination » (Chatterjee, 2004).

6.3 Ouverture à d'autres études et à d'autres méthodologies

6.3.1 Diversification des récits et interculturalité

Questionner l'acceptabilité morale des technologies de transformation de l'émotion dans la voix nécessite d'adresser également l'aspect culturel de cette question et l'importance des narratifs portés consciemment par la société. Comme le soulignaient déjà Rouvroy and Berns (2013) « *dès lors que les machines deviennent de plus en plus « autonomes » et « intelligentes », elles restent bien sûr dépendantes de leur design initial, des intentions, scripts ou scenarii en fonction desquels elles ont été imaginées. Elles sont, dès leur conception (et quelles que soient les formes qu'elles prennent ensuite), porteuses des visions du monde, attentes et projections conscientes ou inconscientes de leurs concepteurs* ».

Nous avons donc initié l'exploration des imaginaires suscités par l'objet filtre vocal chez leurs concepteurs ainsi chez les utilisateurs potentiels, et par la même la question des futurs désirables associés à ces technologies. Ces imaginaires et leurs récits (*narratives* en anglais) sont essentiels au développement scientifique et à l'engagement des individus vis-à-vis des nouvelles technologies. En effet, les récits, qu'ils soient fictifs et non fictifs, ont des effets dans le monde réel. Les narratifs, notamment autour de l'IA peuvent être très utiles, par exemple en inspirant ceux qui travaillent dans les disciplines concernées et dans les secteurs civils, publics et privés, en faisant émerger des futurs alternatifs et en permettant des débats à leur sujet. Néanmoins, ces récits peuvent également créer des attentes et des perceptions erronées qu'il est alors difficile de renverser. Pour quiconque n'est pas familiarisé avec la science ou la technologie, les narratifs peuvent influer sur la perception de leurs applications potentielles et de ceux qui les développent, les promeuvent ou s'y opposent, ainsi que sur le degré de confiance que la société leur accorde. C'est pourquoi ils sont l'objet d'études minutieuses telles que *The AI narratives project* porté par le Centre Leverhulme pour le Futur de l'Intelligence et la Royal Society qui ont entrepris l'examen de la manière dont les chercheurs, décideurs politiques, média et le public parlent de l'IA. Leur rapport (Cave et al., 2018) souligne notamment les limitations des discours actuels qui semblent partager « *une tendance à des extrêmes*

utopistes ou dystopiques et un manque de diversité dans la description des créateurs, protagonistes et types d'IA », contribuant ainsi à des débats mal informés au sujet des technologies issues de l'IA. Dès lors, parmi les recommandations du rapport figure la nécessité d'élargir et surtout diversifier le corpus des récits disponibles.

À la suite de notre étude sur l'acceptabilité morale des techniques de transformations de l'émotion dans la voix, j'ai ainsi réalisé un séjour de recherche de deux mois (sept.-nov. 2022) au sein du laboratoire du Pr Katsumi Watanabe (Department of Intermedia Art and Science, Faculty of Science and Engineering, Waseda University) à Tokyo, financé par une Summer Fellowship de la *Japanese Society for the Promotion of Science* (JSPS), afin d'étudier l'influence de la culture sur les jugements moraux au sujet de l'utilisation des technologies de transformation de la voix. Pour commencer cette étude, j'ai souhaité m'intéresser aux narratifs japonais autour des technologies de transformation de soi. A cette fin, un entretien standardisé a alors été conçu puis mené auprès de divers chercheurs et artistes [1] travaillant sur les questions de l'IA, de la vie artificielle, des robots, et des interactions homme-machine. Chaque rencontre a permis de visiter le laboratoire des chercheurs/ses, de rencontrer l'équipe et de découvrir certains dispositifs (comme le *Face Transformation Mirror* du Dr Yoshida illustré en Figure 6.1 (Yoshida et al., 2013).

Ces entretiens d'une durée moyenne de 2h30 ont été filmés et enregistrés, et restent aujourd'hui à être analysés. L'objectif de ce travail, en plus des données issues du corpus, est de pouvoir présenter ces entretiens aux chercheurs européens afin d'observer puis questionner nos points de convergence ou divergence avec le point de vue japonais, dans le but de vivifier notre réflexion sur cette question sociétale majeure. Il nous semble que cette entreprise pourrait contribuer à la diversification des récits autour des techniques de transformation de voix.

1. Pr Takashi Ikegami, *Graduate School des Arts et des Sciences, The University of Tokyo* – Pr Philippe Codognet, *Laboratoire franco-japonais d'informatique (UMI3527, JFLI)* – Pr Koichiro Eto, *Institut national des sciences et technologies industrielles avancées, Tsukuba* – Pr Yuko Yotsumoto, *Graduate School des Arts et des Sciences, The University of Tokyo* – Pr Olaf Witkowsky, *Cross Labs (Kyoto)*, un institut de recherche sur l'IA, les sciences cognitives et la vie artificielle – Dr Shigeo Yoshida, *OMRON SINIC × Corporation, Tokyo* – Elena Knox, *media artiste et chercheuse*. Ses œuvres questionnent diverses questions dont celle du genre et des relations homme-machine.

FIGURE 6.1 – Dispositif *Face Transformation Mirror* : Miroir déformant l'image du participant pour lui donner l'impression de sourire (Yoshida et al., 2013)

6.3.2 Design fiction

Dans un second temps, à l'occasion des portes ouvertes de l'Ircam à Paris en janvier 2023, nous nous sommes penchés sur la question de l'imaginaire collectif du point de vue des utilisateurs. Pour cela, nous avons conçu avec mon collègue Yann Teytaut de l'équipe Analyse-Synthèse (Laboratoire STMS, CNRS, Paris) un atelier de *Design Fiction* sur le thème des transformations algorithmiques de la voix. Au cours de cet atelier réalisé auprès de 15 volontaires, après une brève introduction aux méthodes de synthèse vocale et d'expérimentation en sciences cognitives, nous avons demandé aux participants de concevoir deux histoires, scénarios fictifs d'utilisation de technologies de transformation de voix, sans se limiter aux possibilités techniques actuelles. Une histoire devait narrer un usage redouté de cette technologie et l'autre un futur désiré permis par l'outil. Une fois ces récits écrits, les participants les racontaient à voix haute au micro ce qui nous donnait l'occasion de leur faire une démonstration par la pratique des techniques d'extraction de paramètres acoustiques (analyses de pitch faites sur les enregistrements) et de les sensibiliser à la méthode scientifique et aux biais d'interprétation qui l'accompagnent. Cela précisé, l'idée principale de cette proposition était alors d'explorer l'imaginaire collectif quant aux potentialités de changements que ces technologies charrient avec elles pour le meilleur et pour le pire. Au sujet du pire, il nous semble qu'adresser la question des peurs de cette manière, à l'inverse de les nourrir, permettrait en leur donner un corps (ou ici une

voix), de les observer avec rigueur et recul et peut être ainsi contribuer à comprendre de quoi elles sont le signe : de craintes légitimes de citoyens au sujet de risques avérés associés au déploiement de ces nouvelles technologies, et/ou de mécompréhensions que des informations correctes pourraient alors clarifier. Le souhait de cet démarche était également celui d'accompagner un discours populaire (et médiatique) parfois anxiogène au sujet de ces nouvelles technologies d'une réflexion orientée non plus uniquement sur ce que l'on craint, mais également sur ce que l'on souhaite voir émerger grâce à ces outils. Cet effort de réorientation du regard et de la pensée en faveur d'un futur désirable, sans tomber dans les excès du transhumanisme, pourrait également contribuer à nourrir autrement un inconscient collectif occidental déjà très empreint de mythes anxiogènes concernant la technologie. A notre sens, cette méthodologie de design fiction que nous avons expérimenté avec beaucoup de plaisir avec Yann Teytaut gagnerait à être répliquée dans un cadre plus formel qui permettrait l'analyse rigoureuse des données qui en ressortiraient.

6.3.3 Acceptabilité morale, le cas des soignants

L'étude éthique présentée au Chapitre 2, qui interrogeait l'acceptabilité morale des filtres vocaux émotionnels dans la population générale, gagnerait il nous semble à être réalisée de manière plus spécifique chez les soignants. Aussi, bien que nous soyons témoins de l'expansion de l'utilisation d'outils de machine learning dans le domaine du soin – en particulier du diagnostique en radiologie et en cancérologie dont les effets bénéfiques semblent être attestés de manière objective (Reddy et al., 2020), nous pouvons nous poser la question de l'accueil qui serait fait à ce genre d'outils au sein de la psychiatrie qui me semble être une spécialité un peu à part, de par la tradition dont elle émerge qui accorde une place particulière à la subjectivité (Kapsambelis, 2018).

Cette question pourrait être adressée par une étude suivant la même méthodologie d'éthique expérimentale que nous présentons dans ce manuscrit mais également au moyen d'ateliers construits sur la base d'entretiens semi-structurés qui permettraient de récolter des informations plus riches à ce sujet. Quelle que soit la méthode, à notre sens, il serait bon de permettre aux participants de tester le dispositif de transformation de voix afin de rendre plus concret l'exercice d'expérience de pensée proposé.

6.3.4 De la perception d'autrui à la perception de soi

Nos études perceptives de l'effet du filtre ont été menées en observant l'effet que la transformation de *pitch* pouvait avoir sur les inférences qu'un individu se fait au sujet d'un interlocuteur. Or, certaines données nous invitent à penser que cet effet pourrait également être observé sur l'utilisateur même du filtre vocal, qu'il s'agisse des informations inférées au sujet de son activité cardiaque ou de son état psychologique dans le cas du TSPT.

Ainsi, nous avons adressé la question de l'influence du filtre vocal dans la perception du RC chez autrui. Or l'observation dans le cas des filtres visuels - objets visant initialement la manipulation de l'image renvoyée à nos interlocuteurs - d'un impact sur la perception de soi (Daar et al., 2021) nous amène à nous poser la question de l'effet du filtre vocal sur l'évaluation de notre propre RC. En effet, au cours des deux dernières décennies, il est devenu largement accepté que notre perception s'étend non seulement à notre environnement extérieur mais aussi aux informations internes à notre corps. La capacité à percevoir les états internes du corps tels que le rythme cardiaque est connue sous le nom d'*interoception*. Ce terme a été proposé pour la première fois par Sherrington (1946) pour la distinguer de l'extéroception (perception de l'environnement extérieur) et de la proprioception (perception des mouvements musculaires)(Tanaka et al., 2021).
La question de l'extension possible de l'effet du filtre à notre interoception est alors soutenue par le cadre théorique de la perception de soi qui implique que *« les individus prennent conscience ou "connaissance" de leurs attitudes, émotions et autres états internes en partie en les inférant à partir de l'observation de leur propres comportements et /ou des circonstances dans lesquels ils se produisent. Selon cette théorie, dans plusieurs circonstances nous agissons avec nous même comme on le ferait avec un observateur extérieur qui devrait se fier à des indices externes pour inférer au sujet d'états internes »* (Bem, 1972).

De surcroît, plusieurs expériences illustrent la possibilité de s'attribuer à tort un objet extérieur, et cela d'autant plus chez les individus à faible conscience intéroceptive. L'expérience du *rubber hand* (Botvinick and Cohen, 1998) est en cela exemplaire de la possibilité de s' « approprier » ou incarner à tort un objet extérieur. Ainsi, en 1998, deux chercheurs de l'université de Pittsburgh ont mené une étude au cours de laquelle les participants étaient assis à une table, un bras caché derrière un écran.

Les chercheurs ont placé un faux bras à sa place, en l'orientant de manière à ce qu'il semble avoir remplacé le vrai bras, puis ont commencé à caresser légèrement la surface des deux bras avec un pinceau. Les participants ont fait état de ce que l'on a appelé l'illusion de la *rubber hand* : ils décrivaient ainsi sentir le pinceau alors même que celui-ci touchait le faux bras. Les individus ayant de faibles capacités intéroceptives semblent particulièrement susceptibles d' « incarner » la *rubber hand*, c'est-à-dire de la percevoir comme leur propre membre, possiblement car ils présenteraient une importante maléabilité de leur représentation corporelle Tsakiris et al. (2011). Cette étude emblématique a été répliquée dans le cadre de la voix par Zheng et al. (2011) qui mettent en évidence une sorte de « rubber voice illusion », dans laquelle la voix d'un étranger, lorsqu'elle est présentée de manière concomitante au discours d'un participant, est perçue comme une version modifiée de sa propre voix (voir aussi Banakou and Slater (2014)).

Toutes ces études amènent à considérer l'intérêt d'un protocole expérimental permettant de tester l'influence qu'un filtre de pitch appliqué sur sa propre voix pourrait avoir sur la perception de son propre état physiologique (RC) et/ou affectif.

De façon spécifique au contexte thérapeutique, l'effet de contagion émotionnelle observé en condition de *vocal feedback* (Aucouturier et al., 2016 ; Goupil et al., 2021 ; Rachman et al., 2017), nous a incité à envisager l'utilisation du filtre de *pitch* (pensé à partir du corpus TraumacoustiK et testé sur la perception des soignants) dans le cadre de la thérapie d'exposition, afin de manipuler en direct la voix du patient alors qu'il narre l'évènement traumatique à l'origine de la pathologie dont il souffre. L'efficacité de la thérapie d'exposition repose en effet en grande partie sur une diminution de l'intensité émotionnelle ressentie par le patient au cours de ces séances (Foa and Kozak, 1986). Le pitch ayant été associé à cette intensité émotionnelle dans notre étude TraumacoustiK, nous considérons l'intérêt du filtre de pitch utilisé en modalité VF comme facilitateur dans le processus de réduction de la charge émotionnelle associée au récit traumatique. A cette fin, nous avons conçu un protocole de recherche clinique qui visera à tester l'effet du filtre de pitch utilisé en VF sur l'efficacité de la thérapie d'exposition en imagination. Le protocole (« Traumavoice », *traitement du trouble de stress post-traumatique par la psychothérapie d'exposition en imagination augmentée par le vocal feedback : étude d'acceptabilité*) a reçu l'avis favorable du

Comité de protection des personnes Île de France XI et permettra nous l'espérons d'approfondir l'étude des effets de l'objet filtre vocal sur nos cognitions.

6.3.5 La transformation de voix : d'un outil méthodologique pour les sciences cognitives à un objet d'étude en soi

D'un point de vue méthodologique enfin, cette thèse se situe dans la continuité du travail réalisé dans le cadre du projet CREAM qui avait pour projet « *d'utiliser la synthèse de sons, et plus exactement l'algorithmie de transformation de signaux sonores, comme technique de manipulation expérimentale des stimuli. Dans cette approche, l'informatique est utilisée pour étendre les possibilités de l'approche expérimentale classique.* » (Aucouturier, 2017) et ainsi renforcer l'arsenal méthodologique à disposition des sciences cognitives.

Cependant, tout en poursuivant l'usage du filtre de pitch comme méthodologie expérimentale (étude de l'influence de l'indice pitch dans la perception au moyen d'une technique de transformation de voix permettant une manipulation causale du système, et non seulement une description corrélationnelle - Casadevall and Fang (2008)), cette thèse amorce également son passage en tant qu'objet d'étude *en soi* : le filtre de voix est en effet aujourd'hui un artefact culturel, dont les effets sur notre cognition peuvent s'étudier pour leur propre valeur, et non une simple technique de laboratoire.

Ce renversement de l'objet d'étude (de la méthode à l'objet), que l'on aurait pu présager de par le potentiel d'augmentation d'une telle technique de transformation de l'expressivité émotionnelle, s'est vu précipité - et cette fois de façon complètement non anticipable - par l'accélération de l'extension du numérique que notre société connaît depuis la pandémie. De fait, cette dernière a grandement popularisé l'usage des outils de visiocommunication dans les domaines à la fois professionnel (réunions zoom, de l'éducation (cours en visio) ou encore du divertissement (concert zoom, apéro WhatsApp etc.). Depuis l'année 2020 qui correspond au début de ma thèse, les scénarios fictifs d'utilisation potentielle de cet outil méthodologique hors du laboratoire se sont transformés en cas d'usage d'un nouvel objet technologique avec lequel il faudra désormais compter, et à sa façon, le glissement méthodologique et conceptuel de ce manuscrit est le témoin de cette période.

6.4 Réflexions personnelles

Tout au long de ce manuscrit, l'utilisation du pronom « nous » a été choisie pour parler des travaux réalisés. Dans cette dernière partie j'adresse des réflexions plus personnelles et écrirai donc à la première personne. J'aborde ici certains questionnements quant à mon positionnement au sujet de ce travail et, de manière plus générale dans ma pratique de chercheuse.

6.4.1 Inscription dans le cadre de la théorie de traitement prédictif

Premièrement, dès le début du manuscrit, je pose mon intention d'observer les effets du filtre vocal à l'intérieur du cadre de la théorie du traitement prédictif. Il me paraît nécessaire ici de préciser que l'adoption de cette théorie me semble relever du pragmatisme dans le sens du concept développé par William James qui ne consisterait pas à présenter une interprétation à l'étude comme étant plus vraie que l'autre, mais s'intéresse aux conséquences pratiques de chacune. De fait, à aucun moment du manuscrit, je ne démontre que ce modèle est le « le bon », l'entreprise de notre travail se situant ailleurs. En outre, examiner notre question en utilisant le modèle de TP me permet de penser et conceptualiser les différents effets du filtre à l'intérieur des neurosciences cognitives en prenant la pleine considération de l'environnement au sein duquel l'objet se déploie. De surcroît, ce cadre théorique présente l'avantage de « *remettre en question l'opposition courante entre la vie biologique, conçue comme pur déterminisme naturel, et la vie symbolique, conçue comme construction culturelle de soi* », attitude prônée par Malabou (2009) qui voit dans ce dépassement « *une méthode pour rendre possible l'élaboration d'une pensée critique au sujet « nouvelles pratiques de transformation de soi » à la hauteur des enjeux philosophiques, scientifiques et technologiques du moment* ».

6.4.2 Choix de la méthode scientifique et naturalisme critique

De plus, au sujet de l'objet filtre vocal, je propose d'en étudier les effets sur nos cognitions sociales fines au moyen de la méthode scientifique expérimentale. Pour autant, je ne considère pas qu'il s'agisse de l'unique méthode valide pour observer cet objet, le regard scientifique n'épuisant pas la chose observée comme le notait Bruno

Latour. Ce dernier, dans la continuité de Simondon (1989) qui décrivait le mode d'existence des objets techniques comme le façonnement d'anthropos par ces objets, parle des modes d'existence des objets afin de rendre compte de la véracité du produit des différentes disciplines (ou modes). Ce projet anthropologique de Latour, critiqué de relativisme pour certains, s'inscrirait selon l'auteur dans le pragmatisme de William James que nous avons abordé. La conscience de la complexité du réel inciterait alors, pour tenter de l'appréhender, à utiliser plus qu'une seule méthode mais une boîte à outils plus fournie.

Au sujet du pragmatisme de James, Bergson (1911) écrivait ainsi : « on comprendrait mal le pragmatisme de James si l'on ne commençait par modifier l'idée qu'on se fait couramment de la réalité en général. On parle du "monde" ou du "cosmos" ; et ces mots, d'après leur origine, désignent quelque chose de simple, tout au moins de bien composé. On dit "l'univers", et le mot fait penser à une unification possible des choses. On peut être spiritualiste, matérialiste, panthéiste, comme on peut être indifférent à la philosophie et satisfait du sens commun : toujours on se représente un ou plusieurs principes simples, par lesquels s'expliquerait l'ensemble des choses matérielles et morales. [...] La réalité, telle que James la voit, est redondante et surabondante. Entre cette réalité et celle que les philosophes reconstruisent, je crois qu'il eût établi le même rapport qu'entre la vie que nous vivons tous les jours et celle que les acteurs nous représentent, le soir, sur la scène. Au théâtre, chacun ne dit que ce qu'il faut dire et ne fait que ce qu'il faut faire [...]. Mais, dans la vie, il se dit une foule de choses inutiles, il se fait une foule de gestes superflus, il n'y a guère de situations nettes ; rien ne se passe aussi simplement, ni aussi complètement, ni aussi joliment que nous le voudrions [...]. Telle est la vie humaine. Et telle est sans doute aussi, aux yeux de James, la réalité en général. »

A mon sens, c'est également la complexité des phénomènes humains qui motiva la création des sciences cognitives (au sein desquelles s'inscrit notre travail) à la croisée de la philosophie, de la psychologie, des neurosciences et de l'informatique. Depuis lors, ces sciences cognitives sont critiquées pour le réductionnisme charrié par leur programme de naturalisation de l'humain. Au sujet du risque d'hypersimplification du réel en vue de le rendre mesurable, il me semble devoir prendre le temps de me pencher sur la thèse du naturalisme critique développée par Daniel Andler au sein de son ouvrage *La silhouette de l'humain* (Andler, 2016), que je découvre seulement

au moment de la rédaction de ce manuscrit. L'auteur y invite à dépasser un rapport dogmatique de complète adhésion ou de pleine méfiance au profit d'un regard critique. Plutôt que de se fonder sur une affirmation naturaliste ontologique ou épistémique concernant la nature de la réalité (naturalisme ontologique : la nature est tout ce qui existe) ou notre capacité à la connaître (naturalisme épistémique : ce que nous pouvons connaître, nous ne pouvons le connaître que de la façon dont nous connaissons la nature), Andler propose un naturalisme principalement « méthodologique » et se concentre sur l'évaluation des programmes de recherche naturalistes par leur capacité à nous fournir une compréhension de la réalité.

Il s'agirait alors, entre autres garde-fous, d'être particulièrement vigilants aux conditions de production des résultats scientifiques. En cela, il me semble qu'il se rapproche du travail de Biezunski and Latour (1989) sur la production du savoir scientifique. Il me faudra creuser cette question plus en profondeur afin de savoir si ce cadre du naturalisme critique est celui qui me permettrait de conjuguer mon souci de la pratique d'une méthodologie scientifique expérimentale et la conscience de l'absence d'hégémonie de ce mode pour décrire le réel.

6.4.3 Positionnement à l'égard des données

La focale mise par Daniel Andler sur l'interprétation des données issues de la méthode scientifique me permet d'aborder un troisième point de réflexion au sujet de mon positionnement dans ce travail : celui à partir duquel j'observe mes données et que j'ai commencé à aborder dans le rationnel de notre étude TraumacoustiK (Chap. 3).

Les conclusions de cette étude m'apparaissent en effet poser en filigrane la question de l'usage des outils du *machine learning* dans le cadre de la démarche diagnostique, en adéquation ou non avec une éthique du soin prenant en considération la singularité des parcours individuels. A mon sens, ces nouveaux outils doivent et peuvent enrichir nos connaissances et perception de la maladie, non l'appauvrir (McQuillan, 2018), mais pour cela il faut répondre au défi épistémologique qu'ils semblent nous imposer. Ainsi, dans un article publié en 2008, Anderson (2008) proclamait « la fin de la théorie » en sciences, la masse des data disponibles permettant à elle seule de rendre compte du réel et rendant ainsi obsolète la méthode scientifique. Ce postulat d'Anderson

procéderait selon Ganascia (2022) du recul de la théorisation associé au risque de dénoétisation scientifique. A rebours de cette tendance, Ganascia (2022) pose que « *le monde brut se donne par l'intermédiaire d'une représentation dans laquelle on le décrit* ». De fait, les données que le chercheur manipule sont issues de mesures crées (qu'il s'agisse d'une expérience en laboratoire ou non) dans le cadre d'un choix théorique fort : elles sont le résultat de ce « parcours d'accès au réel » (Longo, 2023) que Latour avait pris pour objet d'étude anthropologique dans son ouvrage *La science en action* (Biezunski and Latour, 1989). La méthode scientifique suppose la construction humaine du sens, les données seules ne suffisant pas, il s'agira donc d'être particulièrement attentif à l'interprétation que l'on construit. Le soin donné à cette épistémologie contribuerait ainsi au recul nécessaire au bon usage des données corrélationnelles issues des méthodes de *machine learning*. Ganascia souligne d'ailleurs à ce sujet l'utilité pratique de ces données corrélationnelles extraites de manière particulièrement efficaces par ces outils et propose qu'elles « *inspirent à titre heuristique les scientifiques en mettant en évidence telle ou telle hypothèse de travail* », qu'il s'agira ensuite d'aller explorer de manière mécanistique par exemple, une méthodologie que nous avons tenté de suivre dans cette thèse par la manipulation causale de nos variables (pitch et RC).

Face à cela, (Longo, 2023) dans son livre *Le cauchemar de Prométhée* nous renvoie à la nécessité de réinvestir la question du sens afin de pouvoir « maîtriser » pour le mieux nos connaissances scientifiques, et propose pour ce faire que les sciences franchissent la cloison qui les séparent trop souvent de la philosophie. Au cours de ces années de thèse, il m'a semblé que certains concepts empruntés très modestement à philosophie venaient éclairer ma compréhension des effets du filtre vocal et nourrir mes questionnements à ce sujet. Je profite de cette partie plus personnelle pour exprimer l'humilité qui accompagne l'emprunt de ces concepts dont je souhaite pouvoir approfondir la compréhension pendant la suite de ma carrière.

6.4.4 Médiation et médiatisation de la recherche scientifique

Enfin, le dernier point relatif à ma méditation au sujet d'une posture juste concerne la diffusion des connaissances issues de la recherche scientifique hors du cadre académique. Le sujet des deepfakes, de l'IA et des possibilités qu'elle ouvre se prête à

des débats passionnés au sein de la société, en partie peut-être par ce qu'elle pressent le potentiel de mutation que cette technologie charrie avec elle. Dans ce contexte, en tant que chercheuses et chercheurs travaillant sur ces thématiques, nous pouvons alors être interpellés pour venir « éclairer le débat ».

L'idée de rendre la science publique et de partager les connaissances scientifiques n'est pas nouvelle. Selon les travaux de Jurdant (2009), cette idée, en Europe du moins, daterait des Lumières. La nécessaire vulgarisation scientifique est également abordée dans la charte européenne du/de la chercheur-e de 2005 encourageant « *le dialogue social entre les chercheurs et les parties prenantes de la société dans son ensemble* » (Lipani-Vaissade and Pascal, 2020).

Mon sentiment sur ce sujet est qu'il semble nécessaire que les connaissances que nous nous employons à amasser ne circulent pas uniquement à l'intérieur de la sphère de nos pairs. Pour autant, même si je valorise cette diffusion, celle-ci n'en reste pas moins compliquée à plusieurs égards. Ainsi, il convient de noter que jusqu'à présent les scientifiques se livreraient peu à la vulgarisation de leurs travaux. Cette tendance s'expliquerait en partie par le fait que la vulgarisation n'est que faiblement prise en compte par les instances chargées de leur gestion de carrière, comme le souligne Boure (2016). Poupardin and Faury (2018) explicitent ce propos en précisant que la vulgarisation, notamment l'implication des chercheurs dans des contextes non scientifiques informels tels que les débats publics, voir la publication en dehors des cadres formels reconnus par la communauté scientifique, ne « *vaut rien, du point de vue académique si les auteur.e.s ne jouent pas* « le jeu du champ » *et n'acquièrent pas dans le même temps, et au sein de leur discipline d'appartenance, une certaine légitimité* ». En dépit de cette difficulté, la pratique de médiation des travaux scientifiques semble croître, en témoigne notamment la création par les institutions même de recherche, telle que le CNRS d'un média dédié (Le Journal du CNRS). Il reste que ce genre de médium ne bénéficie pas d'une grande audience hors de la sphère universitaire.

Un autre point de complexité à examiner dans cette entreprise de communication scientifique me semble être celui de savoir de quelle parole je suis alors porteuse. En effet, si la communication écrite ou orale à destination du grand public a pu être pour moi un espace d'exploration et de liberté, elle n'en était pas moins un exercice d'équilibriste. Le parti que j'ai choisi de prendre jusqu'à maintenant est celui de présenter les résultats de nos travaux de manière très fidèle à celle faite avec

les co-auteurs et à partir de là, examiner, questionner des implications sociétales de ces résultats qui, quant à elles, pouvaient être le fruit d'une réflexion purement personnelle. Cela a été pour moi l'occasion d'adresser beaucoup plus en profondeur ce type d'argumentation et d'exprimer des réflexions qui ne trouveraient pas leur place au sein de journaux scientifiques de neurosciences ou de psychiatrie qui sont ceux principalement visés pour les travaux issus de cette thèse. En ce qui me concerne, cette médiation est donc d'une grande richesse qui m'a notamment amené à repenser le rapport avec mon objet d'étude et ainsi à amorcer une réflexion critique sur la manière de l'appréhender.

A côté de cela, et de manière souvent intriquée, se pose aussi la question de la médiatisation de la science, un sujet qui n'est pas nouveau non plus mais qui m'a semblé mériter que je m'y penche étant donné la couverture médiatique accordée aux deepfakes et tout ce qui y touche. A mon sens, ici se joue alors une difficulté peut-être plus spécifique qui a fait naître chez moi le souci d'une expression la plus juste possible. *Êtes vous pour ou contre l'utilisation des deepfakes ? Pensez-vous qu'il faille interdire l'usage des deepfakes ? Faut-il ressusciter les morts ? (en référence au deepfake de Dalida ou à la reconstitution de la voix du général de Gaulle). Les deepfakes vont-ils anéantir toute idée de confiance entre les individus ? Faut-il avoir peur des deepfakes ?*. Les incitations à prendre parti sont légion et il est parfois ardu de s'exprimer lorsque l'on peut se sentir sommée de répondre sur un registre émotionnel caractéristique de notre époque selon Quéré (2021) qui souligne que « les faits objectifs ont moins d'influence pour modeler l'opinion publique que les appels à l'émotion et aux opinions personnelles ».

Outre le fait que je ne sois personnellement pas partisane de la capture du sujet par des interventions qui relèvent trop souvent d'idéologies technophobes ou (technoptimistes au choix), il me semble important en tant que chercheuse de tenter de maintenir une expression respectueuse de mon mode qui est celui de la science. Toujours est-il qu'être scientifique n'immunise pas contre les interprétations spontanées et points de vue purement subjectifs. Bachelard (2004) dans *La formation de l'esprit scientifique* le soulignait alors en écrivant que faire de la science consiste à « *penser contre son cerveau* ». Dès lors, afin de tenter de nous prémunir de cette dérive de l'opinion, Etienne Klein propose, au sein de son essai *Le Goût du Vrai* (Klein, 2020), plusieurs points qui pourraient mériter notre attention/prudence. Celui tout d'abord

de limiter son expression à son champs d'expertise, la parole scientifique n'ayant pas réponse à tout. En cela, il m'apparaît indispensable de pouvoir dire sans trop rougir que l'on ne se sent pas habilité à répondre à certaines questions et ainsi à renvoyer vers un collègue plus expert plutôt que de parler avec assurance d'un sujet que l'on ne connaît pas. Peut-être également, serait-il bon d'insister sur le fait que le temps de la recherche est long, plus long que celui de la couverture médiatique et qu'en conséquence il est des questions auxquelles on ne peut pas répondre pour l'instant car les données sont soit inexistantes, soit contradictoires. Enfin, pouvoir argumenter au moyen de résultats scientifiques qui correctement exposés - comment susciter l'intérêt tout en restant fidèle aux données est une question de taille - peuvent être captivants et considérablement apporter aux débats actuels au sujet des usages des technologies issues des études dans le champ de l'IA.

Je tiens à préciser que les « critiques » et points de tension éthique ici soulevés me concernent en premier lieu, il ne s'agit aucunement de donner des leçons mais plutôt de témoigner d'un souci et un questionnement qui ont marqués mes années de doctorat et qui sait, peut-être initier une discussion à ce sujet. D'autant qu'il me semble que face à l'essor des nouvelles technologies issues de l'IA et aux imaginaires particulièrement forts qui leur sont associés, la recherche scientifique pourrait jouer un rôle majeur, celui d'être un agent disruptif (comme le propose par ailleurs Barrau (2023) au sujet de l'écologie), permettant par l'émergence de questions originales de créer un espace de dé-coïncidence fructueux pour la pensée et ainsi de sortir d'une logique du commentaire « contre », de l'exercice de l'indignation qui, lorsqu'il occupe à lui seul un champ trop large ne me semble pas salutaire car ne permettrait pas d'entrevoir des possibles souhaitables auxquels s'atteler. Cette dé-coïncidence est un néologisme amené par Jullien (2017) qui renvoie à un concept opérationnel par lequel le philosophe appelle à défaire les coïncidences idéologiques qui sont pour certaines si collectivement assimilées qu'elles ne sont plus réfléchies, interrogées, et en viendraient alors à bloquer la pensée et figer les débats de société. Comment alors procéder autrement et « débloquer ce qui est en train de s'y figer stérilement » ? Parmi les pistes de réponses, Jullien propose de créer des fissures dans ces coïncidences installées en s'inspirant du modèle de l'artiste : l'artiste n'est tel qu'autant qu'il « décoïncide » de l'art qu'il a déjà fait ou qui est considéré comme tel au temps présent. De même, ce concept s'appliquerait à la pensée. Penser selon Jullien, c'est « décoïncider » de ce qui

s'est déjà pensé, s'écarter de la pensée qui s'est déjà installée ou de ce que « moi-même » j'ai déjà pensé. D'où ma conviction intime que la recherche, par les questions qu'elle pose, aurait les moyens de contribuer à nourrir la réflexion actuelle au sujet de ces nouveaux objets de transformation du soi afin de faire donc dé-coïncider de fausses évidences, pour essayer d'échapper à ce grand ressassement des idées dont nous sommes témoins à ce sujet et qui étriquent la vie de l'esprit.

C'est dans ce contexte de réflexion qu'il m'a semblé nécessaire de mener les présents travaux de recherche sur le potentiel de façonnement cognitif et social de ces nouvelles technologies du soi, avec la conviction que des résultats de ce type contribueront à l'avenir à problématiser les enjeux éthiques qui sont associés à leurs usages et permettront d'éclairer les réflexions au sujet de leur régulation, eu égard aux valeurs jugées souhaitables dans un cadre démocratique telles que l'équité ou la dignité des êtres vivants.

Références

Abe, M., Nakamura, S., Shikano, K., and Kuwabara, H. (1990). Voice conversion through vector quantization. *Journal of the Acoustical Society of Japan (E)*, 11(2) :71–76.

Allard, L. (2009). Pragmatique de l'internet mobile : Technologies de soi et culture du transfert. *Technologies numériques du soi et (co-) constructions identitaires, Paris, L'Harmattan*, pages 59–81.

American Psychiatric Association, D., Association, A. P., et al. (2013). *Diagnostic and statistical manual of mental disorders : DSM-5*, volume 5. American psychiatric association Washington, DC.

Anderson, C. (2008).

Andler, D. (2016). *La silhouette de l'humain. Quelle place pour le naturalisme dans le monde d'aujourd'hui ?* Paris, Gallimard, coll. « NRF Essais ».

Andler, D. (2021). Technologies émergentes et sagesse collective. comprendre, faire comprendre, maîtriser. un vaste programme de plus ? *Les cahiers de Tesaco*.

Aponte-Becerra, L. and Novak, P. (2021). Tilt test : a review. *Journal of Clinical Neurophysiology*, 38(4) :279–286.

Arapova, M. A. (2016). A cross-cultural study of the smile in the russian-and english-speaking world. *Journal of Language and Cultural Education*, 4(2) :56–72.

Arias, P., Belin, P., and Aucouturier, J.-J. (2018a). Auditory smiles trigger unconscious facial imitation. *Current Biology*, 28(14) :R782–R783.

Arias, P., Rachman, L., Liuni, M., and Aucouturier, J.-J. (2020). Beyond correlation : acoustic transformation methods for the experimental study of emotional voice and speech. *Emotion Review*, page 1754073920934544.

Arias, P., Rachman, L., Liuni, M., and Aucouturier, J.-J. (2021). Beyond correlation : Acoustic transformation methods for the experimental study of emotional voice and speech. *Emotion Review*, 13(1) :12–24.

Arias, P., Soladie, C., Bouafif, O., Robel, A., Seguier, R., and Aucouturier, J.-J. (2018b). Realistic transformation of facial and vocal smiles in real-time audiovisual streams. *IEEE Transactions on Affective Computing*.

Arslanova, I., Galvez-Pol, A., Kilner, J., Finotti, G., and Tsakiris, M. (2022). Seeing through each other's hearts : Inferring others' heart rate as a function of own heart rate perception and perceived social intelligence. *Affective Science*, 3(4) :862–877.

ASA (1994). *American National Standard : Acoustical Terminology*. Standards Secretariat, Acoustical Society of America.

Aucouturier, J.-J. (2017). *L'apport des sciences et technologies du son à la recherche en sciences cognitives*. PhD thesis, Université Pierre et Marie Curie.

Aucouturier, J.-J., Johansson, P., Hall, L., Segnini, R., Mercadié, L., and Watanabe, K. (2016). Covert digital manipulation of vocal emotion alter speakers' emotional states in a congruent direction. *Proceedings of the National Academy of Sciences*, 113(4) :948–953.

Ayres, J. G. and Gabbott, P. L. (2002). Vocal cord dysfunction and laryngeal hyperresponsiveness : a function of altered autonomic balance ? *Thorax*, 57(4) :284–285.

Bachelard, G. (2004). *La formation de l'esprit scientifique : contribution à une psychanalyse de la connaissance*. Vrin.

Bachem, A. (1937). Various types of absolute pitch. *The Journal of the Acoustical Society of America*, 9(2) :146–151.

Bachorowski, J.-A. and Owren, M. J. (1995). Vocal expression of emotion : Acoustic properties of speech are associated with emotional intensity and context. *Psychological science*, 6(4) :219–224.

Badour, C. L., Blonigen, D. M., Boden, M. T., Feldner, M. T., and Bonn-Miller, M. O. (2012). A longitudinal test of the bi-directional relations between avoidance coping and ptsd severity during and after ptsd treatment. *Behaviour research and therapy*, 50(10) :610–616.

Banakou, D. and Slater, M. (2014). Body ownership causes illusory self-attribution of speaking and influences subsequent real speaking. *Proceedings of the National Academy of Sciences*, 111(49) :17678–17683.

Banse, R. and Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3) :614.

Bänziger, T., Grandjean, D., Bernard, P.-J., Klasmeyer, G., and Scherer, K. R. (2001). Prosodie de l'émotion : étude de l'encodage et du décodage. *Cahiers de linguistique française*, 23 :11–37.

Barrau, A. (2023). *L'Hypothèse K. La science face à la catastrophe écologique*. Grasset.

Barrett, L. F. (2012). Emotions are real. *Emotion*, 12(3) :413–429.

Barrows, H. S. and Feltovich, P. J. (1987). The clinical reasoning process. *Medical education*, 21(2) :86–91.

Beauchaine, T. P. (2015). Respiratory sinus arrhythmia : A transdiagnostic biomarker of emotion dysregulation and psychopathology. *Current opinion in psychology*, 3 :43–47.

Beller, G. (2010). Expresso : transformation of expressivity in speech. In *Speech Prosody 2010-Fifth International Conference*.

Bem, D. J. (1972). Self-perception theory. In *Advances in experimental social psychology*, volume 6, pages 1–62. Elsevier.

Ben-Ze'ev, A. (1997). Emotions and morality. *The Journal of Value Inquiry*, 31(2) :195–212.

Bergson, H. (1911). Sur le pragmatisme de william james. vérité et réalité. *JAMES, William. Le pragmatisme. Tradução de E. Le Brun. Paris : Flammarion*.

Biezunski, M. and Latour, B. (1989). La science en action.

Boenink, M., Swierstra, T., and Stemerding, D. (2010). Anticipating the interaction between technology and morality : A scenario study of experimenting with humans in bionanotechnology. *Studies in ethics, law, and technology*, 4(2).

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot. Int.*, 5(9) :341–345.

Boersma P., W. D. (2007). Praat : Doing phonetics by computer, version 5.2.34. *Available online : http ://www.fon.hum.uva.nl/praat/*.

Boidron, L., Boudenia, K., Avena, C., Boucheix, J.-M., and Aucouturier, J.-J. (2016). Emergency medical triage decisions are swayed by computer-manipulated cues of physical dominance in caller's voice. *Scientific reports*, 6(1) :30219.

Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, 352(6293) :1573–1576.

Bottemanne, H. (2021). Cerveau bayésien : peut-on modéliser l'émotion ? *L'Encéphale*, 47(1) :58–63.

Bottemanne, H., Longuet, Y., and Gauld, C. (2022). L'esprit predictif : introduction à la théorie du cerveau bayésien. *L'Encéphale*, 48(4) :436–444.

Botvinick, M. and Cohen, J. (1998). Rubber hands "feel" touch that eyes see. *Nature*, 391(6669) :756.

Boure, R. (2016). Les paroles de chercheurs, le numérique et la scène sociale. *Mondes sociaux*.

Bower, G. H. (1979). *Psychology of learning and motivation*. Academic Press.

Bradley, R., Greene, J., Russ, E., Dutra, L., and Westen, D. (2005). A multidimensional meta-analysis of psychotherapy for ptsd. *American journal of Psychiatry*, 162(2) :214–227.

Breiman, L. (2001). Random forests. *Machine learning*, 45 :5–32.

Breslau, N., Davis, G. C., Andreski, P., and Peterson, E. (1991). Traumatic events and posttraumatic stress disorder in an urban population of young adults. *Archives of general psychiatry*, 48(3) :216–222.

Brewin, C. R. (2006). Understanding cognitive behaviour therapy : A retrieval competition account. *Behaviour research and therapy*, 44(6) :765–784.

Brewin, C. R. (2018). Memory and forgetting. *Current Psychiatry Reports*, 20(10) :87.

Brown, L. A., Zandberg, L. J., and Foa, E. B. (2019). Mechanisms of change in prolonged exposure therapy for ptsd : Implications for clinical practice. *Journal of Psychotherapy Integration*, 29(1) :6–14.

Brunet, A., Ayrolles, A., Gambotti, L., Maatoug, R., Estellat, C., Descamps, M., Girault, N., Kalalou, K., Abgrall, G., Ducrocq, F., et al. (2019). Paris mem : a study protocol for an effectiveness and efficiency trial on the treatment of traumatic stress in france after the 2015–16 terrorist attacks. *BMC psychiatry*, 19 :1–9.

Brunet, A., Poundja, J., Tremblay, J., Bui, É., Thomas, É., Orr, S. P., Azzoug, A., Birmes, P., and Pitman, R. K. (2011). Trauma reactivation under the influence of

propranolol decreases posttraumatic stress symptoms and disorder : 3 open-label trials. *Journal of clinical psychopharmacology*, 31(4) :547–550.

Brunet, A., Saumier, D., Liu, A., Streiner, D. L., Tremblay, J., and Pitman, R. K. (2018). Reduction of ptsd symptoms with pre-reactivation propranolol therapy : a randomized controlled trial. *American Journal of Psychiatry*, 175(5) :427–433.

Burred, J. J., Ponsot, E., Goupil, L., Liuni, M., and Aucouturier, J.-J. (2019). Cleese : An open-source audio-transformation toolbox for data-driven experiments in speech and music cognition. *PloS one*, 14(4) :e0205943.

Cabrera, L. Y., Fitz, N. S., and Reiner, P. B. (2015). Empirical support for the moral salience of the therapy-enhancement distinction in the debate over cognitive, affective and social enhancement. *Neuroethics*, 8(3) :243–256.

Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., and Snyder, P. J. (2004). Voice acoustical measurement of the severity of major depression. *Brain and cognition*, 56(1) :30–35.

Casadevall, A. and Fang, F. C. (2008). Descriptive science. *Infection and immunity*, 76(9) :3835.

Cave, S., Craig, C., Dihal, K., Dillon, S., Montgomery, J., Singler, B., and Taylor, L. (2018). *Portrayals and perceptions of AI and why they matter*.

CCNE (2013).

Charlin, B., Boshuizen, H. P., Custers, E. J., and Feltovich, P. J. (2007). Scripts and clinical reasoning. *Medical education*, 41(12) :1178–1184.

Charlin, B., Tardif, J., and Boshuizen, H. P. (2000). Scripts and medical diagnostic knowledge : theory and applications for clinical reasoning instruction and research. *Academic medicine*, 75(2) :182–190.

Chatterjee, A. (2004). Cosmetic neurology : the controversy over enhancing movement, mentation, and mood. *Neurology*, 63(6) :968–974.

Chong, C. S., Kim, J., and Davis, C. (2018). Disgust expressive speech : The acoustic consequences of the facial expression of emotion. *Speech Communication*, 98 :68–72.

Clark, A. (2015). *Surfing uncertainty : Prediction, action, and the embodied mind*. Oxford University Press.

Clark, A. (2023). The experience machine : how our minds predict and shape reality.

Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis*, 58(1) :7–19.

Clarke, A. E., Shim, J. K., Mamo, L., Fosket, J. R., and Fishman, J. R. (2003). Biomedicalization : Technoscientific transformations of health, illness, and us biomedicine. *American sociological review*, pages 161–194.

Costa, J., Jung, M. F., Czerwinski, M., Guimbretière, F., Le, T., and Choudhury, T. (2018). Regulating feelings during interpersonal conflicts by changing voice self-perception. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 1–13, Montreal QC Canada. ACM.

Cristel, R. T., Dayan, S. H., Akinosun, M., and Russell, P. T. (2021). Evaluation of selfies and filtered selfies and effects on first impressions. *Aesthetic Surgery Journal*, 41(1) :122–130.

Crocq, M.-A., Guelfi, J.-D., Boyer, P., Pull, C.-B., Marie-Claire, P., Association, A. P., et al. (2016). *Mini DSM-5 Critères Diagnostiques*. Elsevier Health Sciences.

Daar, D. A., Chiodo, M. V., and Rohrich, R. J. (2021). The zoom view : How does video conferencing affect what our patients see in themselves, and how can we do right by them ? *Plastic and Reconstructive Surgery*, 148(1) :172–174.

Darwin, C. (1872). *The expression of the emotions in man and animals*. 1998 Ed. : Oxford University Press, USA.

Dattorro, J. (1997). Effect design, part 2 : Delay line modulation and chorus. *Journal of the Audio engineering Society*, 45(10) :764–788.

Davis, N. C. (2007). Smile design. *Dental Clinics of North America*, 51(2) :299–318.

de Leeuw, J. R., Gilbert, R. A., and Luchterhandt, B. (2023). jspsych : Enabling an open-source collaborative ecosystem of behavioral experiments. *Journal of Open Source Software*, 8(85) :5351.

Degottex, G., Lanchantin, P., Roebel, A., and Rodet, X. (2013). Mixed source model and its adapted vocal tract filter estimate for voice transformation and synthesis. *Speech Communication*, 55(2) :278–294.

Degottex, G., Roebel, A., and Rodet, X. (2011). Pitch transposition and breathiness modification using a glottal source model and its adapted vocal-tract filter. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5128–5131. IEEE.

Dubois, J., Giacomo, M., Guespin, L., Marcellesi, C., Marcellesi, J.-B., and Mével, J.-P. (1974). Dictionnaire de linguistique. *(No Title)*.

Ekkekakis, P. (2013). *The measurement of affect, mood, and emotion : A guide for health-behavioral research*. Cambridge University Press.

El-Hage, W., Birmes, P., Jehel, L., Ferreri, F., Benoit, M., Vidailhet, P., Prieto, N., François, I., Baubet, T., and Vaiva, G. (2019). Improving the mental health system for trauma victims in france. *European journal of psychotraumatology*, 10(1) :1617610.

Elias, J. J., Lacetera, N., and Macis, M. (2019). Paying for kidneys ? a randomized survey and choice experiment. *American Economic Review*, 109(8) :2855–88.

Ellgring, H. and Scherer, K. R. (1996). Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior*, 20 :83–110.

Ellsworth, P. C. (1994). William james and emotion : is a century of fame worth a century of misunderstanding ? *Psychological review*, 101(2) :222.

Elstein, A. S. (1999). Heuristics and biases : selected errors in clinical reasoning. *Academic Medicine*, 74(7) :791–4.

Elstein, A. S., Shulman, L. S., and Sprafka, S. A. (1978). *Medical problem solving : An analysis of clinical reasoning*. Harvard University Press.

Epstein, R. M. (1999). Mindful practice. *Jama*, 282(9) :833–839.

Eva, K. W. (2005). What every teacher needs to know about clinical reasoning. *Medical education*, 39(1) :98–106.

Eva, K. W., Hatala, R. M., LeBlanc, V. R., and Brooks, L. R. (2007). Teaching from the clinical reasoning literature : combined reasoning strategies help novice diagnosticians overcome misleading information. *Medical education*, 41(12) :1152–1158.

Farner, S., Röbel, A., and Rodet, X. (2009). Natural transformation of type and nature of the voice for extending vocal repertoire in high-fidelity applications. In *Audio Engineering Society Conference : 35th International Conference : Audio for Games*. Audio Engineering Society.

Farrús, M., Hernando, J., and Ejarque, P. (2007). Jitter and shimmer measurements for speaker recognition. In *8th Annual Conference of the International Speech Communication Association ; 2007 Aug. 27-31 ; Antwerp (Belgium).[place unknown] : ISCA ; 2007. p. 778-81*. International Speech Communication Association (ISCA).

Fenster, R. J., Lebois, L. A. M., Ressler, K. J., and Suh, J. (2018). Brain circuit dysfunction in post-traumatic stress disorder : from mouse to man. *Nature reviews. Neuroscience*, 19(9) :535–551.

Foa, E. B. and Kozak, M. J. (1986). Emotional processing of fear : exposure to corrective information. *Psychological bulletin*, 99(1) :20.

Foa, E. B. and Rothbaum, B. O. (1998). *Treating the trauma of rape : Cognitive-behavioral therapy for PTSD*. Guilford Press.

Foa, E. B., Rothbaum, B. O., Riggs, D. S., and Murdock, T. B. (1991). Treatment of posttraumatic stress disorder in rape victims : a comparison between cognitive-behavioral procedures and counseling. *Journal of consulting and clinical psychology*, 59(5) :715.

for Parliamentary Research Services., E. P. D. G. (2019). *Polarisation and the use of technology in political campaigns and communication.* Publications Office, LU.

Foucault, M. (1983). Usage des plaisirs et techniques de soi. *Le Débat*, (5) :46–72.

Frahm, J. (2018). Beatboxing in real time.

Friedman, M. J. (2018). Eradicating traumatic memories : Implications for ptsd treatment. *American Journal of Psychiatry*, 175(5) :391–392.

Friston, K. and Frith, C. (2015a). A duet for one. *Consciousness and cognition*, 36 :390–405.

Friston, K. J. and Frith, C. D. (2015b). Active inference, communication and hermeneutics. *Cortex*, 68 :129–143.

Galvez-Pol, A., Antoine, S., Li, C., and Kilner, J. M. (2022). People can identify the likely owner of heartbeats by looking at individuals' faces. *Cortex*, 151 :176–187.

Ganascia, J.-G. (2022). Enjeux épistémologiques de la science des données. In *Annales des Mines-Réalités industrielles*, number 3, pages 45–48. Cairn/Softwin.

Gendron, M. and Barrett, L. F. (2018). Emotion perception as conceptual synchrony. *Emotion Review*, 10(2) :101–110.

Gillie, B. L. and Thayer, J. F. (2014). Individual differences in resting heart rate variability and cognitive control in posttraumatic stress disorder. 5.

Godin, K. W. and Hansen, J. H. (2015). Physical task stress and speaker variability in voice quality. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015 :1–13.

Godulla, A., Hoffmann, C., and Seibert, D. (2021). Dealing with deepfakes - an interdisciplinary examination of the state of research and implications for communication studies. *Studies in Communication and Media*, 10.

Goffette, J. (2006). *Naissance de l'anthropotechnie : de la médecine au modelage de l'humain*. Vrin.

Goffi, J.-Y. (2009). *Thérapie, augmentation et finalité de la médecine*. na.

Goffman, E. (1973). La mise en scène de la vie quotidienne, t. 1 la présentation de soi, trad. *A. Accardo, Paris, Les Editions de Minuit*.

Goupil, L., Ponsot, E., Richardson, D., Reyes, G., and Aucouturier, J.-J. (2021). Listeners' perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nature communications*, 12(1) :1–17.

GOV.UK (2019). Snapshot paper - smart speakers and voice assistants. *GOV UK*.

Graham, J., Nosek, B. A., Haidt, J., Iyer, R., Koleva, S., and Ditto, P. H. (2011). Mapping the moral domain. *Journal of personality and social psychology*, 101(2) :366.

Gross, J. J. and Jazaieri, H. (2014). Emotion, emotion regulation, and psychopathology : An affective science perspective. *Clinical psychological science*, 2(4) :387–401.

Guerouaou, N., Vaiva, G., and Aucouturier, J.-J. (2021). The shallow of your smile : the ethics of expressive vocal deep-fakes. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 377(1841) :20210083.

Gurfinkel, J. (2023). Ai and the american smile. [Online ; posted 27-March-2023].

Hancock, J. T., Naaman, M., and Levy, K. (2020). Ai-mediated communication : Definition, research agenda, and ethical considerations. *Journal of Computer-Mediated Communication*, 25(1) :89–100.

Haraway, D. (1984). Manifeste cyborg : Science, technologie et féminisme socialiste à la fin du xxe siècle. *Mouvements*, 45–46(3–4) :15–21.

Heidegger, M. (1958). *La question de la technique*, volume 52. Gallimard Paris.

Henrich, J., Heine, S. J., and Norenzayan, A. (2010). Most people are not weird. *Nature*, 466(7302) :29–29.

Hochschild, A. R., Fournet-Fayas, S., and Thome, C. (2017). *Le prix des sentiments : au cœur du travail émotionnel*. La découverte.

Hopper, J. W., Frewen, P. A., Sack, M., Lanius, R. A., and Van Der Kolk, B. A. (2007). The responses to script-driven imagery scale (rsdi) : Assessment of state post-

traumatic symptoms for psychobiological and treatment research. *Journal of Psychopathology and Behavioral Assessment*, 29(4) :249–268.

Ilie, G. and Thompson, W. F. (2006). A comparison of acoustic cues in music and speech for three dimensions of affect. *Music Perception*, 23(4) :319–330.

Illouz, E. (2006). *Les sentiments du capitalisme*. Seuil Paris.

Jack, R. E., Garrod, O. G., Yu, H., Caldara, R., and Schyns, P. G. (2012). Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19) :7241–7244.

Jack, R. E., Sun, W., Delis, I., Garrod, O. G., and Schyns, P. G. (2016). Four not six : Revealing culturally common facial expressions of emotion. *Journal of Experimental Psychology : General*, 145(6) :708.

Jacobs, B. (2018). Techniques de soi et technologies numériques. *Entrelacs. Cinéma et audiovisuel*, (1515).

Jakesch, M., Hancock, J. T., and Naaman, M. (2023). Human heuristics for ai-generated language are flawed. *Proceedings of the National Academy of Sciences*, 120(11) :e2208839120.

Javornik, A., Marder, B., Barhorst, J. B., McLean, G., Rogers, Y., Marshall, P., and Warlop, L. (2022). 'what lies behind the filter ?' uncovering the motivations for using augmented reality (ar) face filters on social media and their effect on well-being. *Computers in Human Behavior*, 128 :107126.

Jiang, X. and Pell, M. D. (2017). The sound of confidence and doubt. *Speech Communication*, 88 :106–126.

Jullien, F. (2017). *Dé-coïncidence : D'où viennent l'art et l'existence ?* Grasset.

Jurdant, B. (2009). *Les problèmes théoriques de la vulgarisation scientifique*. Archives contemporaines.

Juslin, P. N. and Laukka, P. (2003). Communication of emotions in vocal expression and music performance : Different channels, same code ? *Psychological bulletin*, 129(5) :770.

Juslin, P. N. and Västfjäll, D. (2008). Emotional responses to music : The need to consider underlying mechanisms. *Behavioral and brain sciences*, 31(5) :559–575.

Kahane, G. (2015). Sidetracked by trolleys : Why sacrificial moral dilemmas tell us little (or nothing) about utilitarian judgment. *Social neuroscience*, 10(5) :551–560.

Kapsambelis, V. (2018). *Manuel de psychiatrie clinique et psychopathologique de l'adulte*. PR DE L'UNIV DU QUEBEC.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119.

Kass, L. (2003). *Beyond therapy : biotechnology and the pursuit of happiness*. Executive Office of the President.

Kawahara, H. (1997). Speech representation and transformation using adaptive interpolation of weighted spectrum : vocoder revisited. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages 1303–1306. IEEE.

Khosravani, S., Mahnan, A., Yeh, I.-L., Aman, J. E., Watson, P. J., Zhang, Y., Goding, G., and Konczak, J. (2019). Laryngeal vibration as a non-invasive neuromodulation therapy for spasmodic dysphonia. *Scientific reports*, 9(1) :1–11.

Kindt, M. (2018). The surprising subtleties of changing fear memory : a challenge for translational science. *Philosophical Transactions of the Royal Society B : Biological Sciences*, 373(1742) :20170033.

Klein, É. (2020). *Le goût du vrai*. Gallimard.

Knapp, M. L., Hall, J. A., and Horgan, T. G. (2013). *Nonverbal communication in human interaction*. Cengage Learning.

Körner, A., Deutsch, R., and Gawronski, B. (2020). Using the cni model to investigate individual differences in moral dilemma judgments. *Personality and Social Psychology Bulletin*, 46(9) :1392–1407.

Koverola, M., Drosinou, M., Palomäki, J., Halonen, J., Kunnari, A., Repo, M., Lehtonen, N., and Laakasuo, M. (2020a). Moral psychology of sex robots : An experimental study- how pathogen disgust is associated with interhuman sex but not interandroid sex. *Paladyn, Journal of Behavioral Robotics*, 11(1) :233–249.

Koverola, M., Kunnari, A., Drosinou, M., Palomäki, J., Hannikainen, I. R., Sundvall, J., and Laakasuo, M. (2020b). Non-human superhumans-moral psychology of brain implants : Exploring the role of situational factors, science fiction exposure, individual differences and perceived norms.

Kreiman, J. and Gerratt, B. R. (2005). Perception of aperiodicity in pathological voice. *The Journal of the Acoustical Society of America*, 117(4) :2201–2211.

Kundinger, T., Sofra, N., and Riener, A. (2020). Assessment of the potential of wrist-worn wearable sensors for driver drowsiness detection. *Sensors*, 20(4) :1029.

Laakasuo, M., Drosinou, M., Koverola, M., Kunnari, A., Halonen, J., Lehtonen, N., and Palomäki, J. (2018). What makes people approve or condemn mind upload technology ? untangling the effects of sexual disgust, purity and science fiction familiarity. *Palgrave Communications*, 4(1) :1–14.

Laakasuo, M., Palomäki, J., and Köbis, N. (2021). Moral uncanny valley : A robot's appearance moderates how its decisions are judged. *International Journal of Social Robotics*, 13(7) :1679–1688.

Lang, P. J. (1977). Imagery in therapy : An information processing analysis of fear. *Behavior therapy*, 8(5) :862–886.

Lange, K., Kühn, S., and Filevich, E. (2015). " just another tool for online studies"(jatos) : An easy solution for setup and management of web servers supporting online studies. *PloS one*, 10(6) :e0130834.

Lanius, R., Bluhm, R., Lanius, U., and Pain, C. (2006). A review of neuroimaging studies in ptsd : heterogeneity of response to symptom provocation. *Journal of psychiatric research*, 40(8) :709–729.

Laxar, K., Luria, S., and CT, N. S. M. R. L. G. (1990). Frequency of a flashing light as a navigational range indicator. *NSMRL Report*, 1157.

LeDoux, J. E. and Hofmann, S. G. (2018). The subjective experience of emotion : a fearful view. *Current Opinion in Behavioral Sciences*, 19 :67–72.

Leightley, D., Williamson, V., Darby, J., and Fear, N. T. (2019). Identifying probable post-traumatic stress disorder : applying supervised machine learning to data from a uk military cohort. *Journal of Mental Health*, 28(1) :34–41.

Lennerfors, T. T. and Murata, K. (2019). *Tetsugaku Companion to Japanese Ethics and Technology*. Springer Verlag.

Leroy, A., Very, E., Birmes, P., Yger, P., Szaffarczyk, S., Lopes, R., Outteryck, O., Faure, C., Duhem, S., Grandgenevre, P., et al. (2022). Intrusive experiences in post-traumatic stress disorder : treatment response induces changes in the directed functional connectivity of the anterior insula. *NeuroImage : Clinical*, 34 :102964.

Levenson, R. W. (2014). The autonomic nervous system and emotion. *Emotion Review*, 6(2) :100–112.

Lipani-Vaissade, M.-C. and Pascal, C. (2020). Recherche scientifique et médias : enjeux et tensions. *Revue française des Sciences de l'Information, RFSIC*, (20, septembre 2020).

Liuni, M. and Röbel, A. (2013). Phase vocoder and beyond. *Musica/Tecnologia*, pages 73–89.

Lonergan, M., Olivera-Figueroa, L., Pitman, R., and Brunet, A. (2013). Propranolol's effects on the consolidation and reconsolidation of long-term emotional memory in healthy participants : a meta-analysis. *Journal of Psychiatry & Neuroscience*, 38(4) :222–231.

Longo, G. (2023). *Le cauchemar de Prométhée : Les sciences et leurs limites*. PUF.

Luo, Z., Chen, J., Takiguchi, T., and Ariki, Y. (2017). Emotional voice conversion using neural networks with arbitrary scales f0 based on wavelet transform. *EURASIP Journal on Audio, Speech, and Music Processing*, 2017(1) :1–13.

Malabou, C. (2009). *Changer de différence, Le féminin et la question philosophique*. Galilée.

Marmar, C. R., Brown, A. D., Qian, M., Laska, E., Siegel, C., Li, M., Abu-Amara, D., Tsiartas, A., Richey, C., Smith, J., Knoth, B., and Vergyri, D. (2019). Speech-based markers for posttraumatic stress disorder in us veterans. *Depression and Anxiety*, 36(7) :607–616.

Marsac, J. (2013). Variabilité de la fréquence cardiaque : un marqueur de risque cardiométabolique en santé publique. *Bulletin de l'Académie nationale de médecine*, 197(1) :175–186.

Matsubara, M., Augereau, O., Sanches, C. L., and Kise, K. (2016). Emotional arousal estimation while reading comics based on physiological signal analysis. In *Proceedings of the 1st International Workshop on comics ANalysis, Processing and Understanding*, pages 1–4.

Matsumoto, D., Yoo, S. H., and Fontaine, J. (2008). Mapping expressive differences around the world : The relationship between emotional display rules and individualism versus collectivism. *Journal of cross-cultural psychology*, 39(1) :55–74.

McCarthy, C., Pradhan, N., Redpath, C., and Adler, A. (2016). Validation of the empatica e4 wristband. In *2016 IEEE EMBS international student conference (ISC)*, pages 1–4. IEEE.

McLean, C. P., Levy, H. C., Miller, M. L., and Tolin, D. F. (2022). Exposure therapy for ptsd : A meta-analysis. *Clinical psychology review*, 91 :102115.

McQuillan, D. (2018). Mental health and artificial intelligence : losing your-voice. https ://www.opendemocracy.net/en/digitaliberties/mental-health-and-artificial-intelligence-losing-your-voice-poem.

Medaglia, J. D., Yaden, D. B., Helion, C., and Haslam, M. (2019). Moral attitudes and willingness to enhance and repair cognition with brain stimulation. *Brain stimulation*, 12(1) :44–53.

Métayer, S. and Pahlavan, F. (2014). Validation of the moral foundations questionnaire in french. *Revue internationale de psychologie sociale*, 27(2) :79–107.

Michaelis, D., Fröhlich, M., Strube, H. W., Kruse, E., Story, B., and Titze, I. R. (1998). Some simulations concerning jitter and shimmer measurement. In *3rd International Workshop on Advances in Quantitative Laryngoscopy, Aachen, Germany*, pages 744–754.

Michaud, T. (2020). *Science Fiction and Innovation Design*. John Wiley & Sons.

Mohammadi, S. H. and Kain, A. (2017). An overview of voice conversion systems. *Speech Communication*, 88 :65–82.

Moine, C. L. and Obin, N. (2020). Att-hack : An expressive speech database with social attitudes. *arXiv preprint arXiv :2004.04410*.

Moulines, E. and Laroche, J. (1995). Non-parametric techniques for pitch-scale and time-scale modification of speech. *Speech communication*, 16(2) :175–205.

Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K., and Geralts, D. S. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. *Journal of neurolinguistics*, 20(1) :50–64.

Nader, K., Kriegler, K., Blake, D., Pynoos, R., Newman, E., and Weathers, F. (1996). Clinician-administered ptsd scale for children and adolescents.

Newman, S. and Mather, V. G. (1938). Analysis of spoken language of patients with affective disorders. *American journal of psychiatry*, 94(4) :913–942.

Nightingale, S. J. and Farid, H. (2022). Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8) :e2120481119.

Onton, J. A., Kang, D. Y., and Coleman, T. P. (2016). Visualization of whole-night sleep eeg from 2-channel mobile recording device reveals distinct deep sleep stages with differential electrodermal activity. *Frontiers in human neuroscience*, 10 :605.

Orgad, S. and Gill, R. (2021). *Confidence culture*. Duke University Press.

Orlikoff, F. R. and Baken, R. J. (1989). The effect of the heartbit on fondamental vocal frequency perturbation. *Journal of Speech and Hearing Research*, 32 :576–582.

Orlikoff, R. F. (1989). Vocal jitter at different fundamental frequencies : A cardiovascular-neuromuscular explanation. *Journal of Voice*, 3(2) :104–112.

Orr, S. P. and Roth, W. T. (2000). Psychophysiological assessment : Clinical applications for ptsd. *Journal of affective Disorders*, 61(3) :225–240.

O'Donnell, M. L., Elliott, P., Lau, W., and Creamer, M. (2007). Ptsd symptom trajectories : From early to chronic response. *Behaviour Research and Therapy*, 45(3) :601–606.

Pajon, P. (2012). Repenser la nature, dialogue philosophique, europe, asie, amériques. In *Cybersphère et industries anthropotechniques : quelques exemples et questions*, pages 377–392. Presse de l'université de Laval.

Parkinson, B. and Manstead, A. S. R. (1993). Making sense of emotion in stories and social life. *Cognition and Emotion*, 7(3-4) :295–323.

Pelaccia, T., Tardif, J., Triby, E., Ammirati, C., Betrand, C., and Charlin, B. (2011). Comment les médecins raisonnent-ils pour poser des diagnostics et prendre des décisions thérapeutiques ? les enjeux en médecine d'urgence. *Annales françaises de médecine d'urgence*, 1(1) :77–84.

Pell, M. D. and Skorup, V. (2008). Implicit processing of emotional prosody in a foreign versus native language. *Speech Communication*, 50(6) :519–530.

Persson, I. and Savulescu, J. (2008). The perils of cognitive enhancement and the urgent imperative to enhance the moral character of humanity. *Journal of applied philosophy*, 25(3) :162–177.

Pine, D. S. and LeDoux, J. E. (2017). Elevating the role of subjective experience in the clinic : response to fanselow and pennington. *American Journal of Psychiatry*, 174(11) :1121–1122.

Pitman, R. K. (1989). Post-traumatic stress disorder, hormones, and memory. *Biological psychiatry*.

Ponsot, E., Arias, P., and Aucouturier, J. (2018). Uncovering mental representations of smiled speech using reverse correlation. *Journal of the Acoustical Society of America (submitted)*.

Poupardin, E. and Faury, M. (2018). Hypothèses : l'inscription d'une pratique de communication dans l'activité de recherche. *Revue Française des sciences de l'information et de la communication*, (15).

Prieur, C., Dourgnon, P., Jusot, F., Marsaudon, A., Wittwer, J., and Guillaume, S. (2022). Une personne sans titre de séjour sur six souffre de troubles de stress post-traumatique en france. *Questions d'économie de la santé*, (266) :8p.

Primov-Fever, A., Lidor, R., Meckel, Y., and Amir, O. (2014). The effect of physical effort on voice characteristics. *Folia Phoniatrica et Logopaedica*, 65(6) :288–293.

Pruvost-Robieux, E., André-Obadia, N., Marchi, A., Sharshar, T., Liuni, M., Gavaret, M., and Aucouturier, J.-J. (2022). It's not what you say, it's how you say it : A retrospective study of the impact of prosody on own-name p300 in comatose patients. *Clinical Neurophysiology*, 135 :154–161.

Pumarola, A., Agudo, A., Martinez, A. M., Sanfeliu, A., and Moreno-Noguer, F. (2018). Ganimation : Anatomically-aware facial animation from a single image. In *Proceedings of the European conference on computer vision (ECCV)*, pages 818–833.

Puts, D. A., Gaulin, S. J., and Verdolini, K. (2006). Dominance and the evolution of sexual dimorphism in human voice pitch. *Evolution and human behavior*, 27(4) :283–296.

Quéré, L. (2021).

Rachman, L., Liuni, M., Arias, P., Lind, A., Johansson, P., Hall, L., Richardson, D., Watanabe, K., Dubal, S., and Aucouturier, J.-J. (2017). David : An open-source platform for real-time transformation of infra-segmental emotional cues in running speech. *Behavior Research Methods*, pages 1–21.

Reddy, S., Allan, S., Coghlan, S., and Cooper, P. (2020). A governance model for the application of ai in health care. *Journal of the American Medical Informatics Association*, 27(3) :491–497.

Révis, J. (2013). *La voix et soi : Ce que notre voix dit de nous*. De Boeck Superieur.

Robb, L. (1999). Emotional musicality in mother-infant vocal affect, and an acoustic study of postnatal depression. *Musicae Scientiae*, 3(1) :123–154.

Roullet, P., Vaiva, G., Véry, E., Bourcier, A., Yrondi, A., Dupuch, L., Lamy, P., Thalamas, C., Jasse, L., El Hage, W., et al. (2021). Traumatic memory reactivation with or without propranolol for ptsd and comorbid md symptoms : a randomised clinical trial. *Neuropsychopharmacology*, 46(9) :1643–1649.

Rouvroy, A. and Berns, T. (2013). Gouvernementalité algorithmique et perspectives d'émancipation. *Réseaux*, 177(1) :163–196.

Rychlowska, M., Miyamoto, Y., Matsumoto, D., Hess, U., Gilboa-Schechtman, E., Kamble, S., Muluk, H., Masuda, T., and Niedenthal, P. M. (2015). Heterogeneity of long-history migration explains cultural differences in reports of emotional expressivity and the functions of smiles. *Proceedings of the National Academy of Sciences*, 112(19) :E2429–E2436.

Sacharin, V., Schlegel, K., and Scherer, K. R. (2012). Geneva emotion wheel rating study. *Center for Person, Kommunikation, Aalborg University, NCCR Affective Sciences. Aalborg University, Aalborg*.

Sack, M., Cillien, M., and Hopper, J. W. (2012). Acute dissociation and cardiac reactivity to script-driven imagery in trauma-related disorders. *European Journal of Psychotraumatology*, 3(1) :17419.

Sadin, É. (2018). *L'intelligence artificielle*.

Safra, L., Chevallier, C., Grèzes, J., and Baumard, N. (2020). Tracking historical changes in trustworthiness using machine learning analyses of facial cues in paintings. *Nature communications*, 11(1) :1–7.

Sahakian, B. J. and Morein-Zamir, S. (2011). Neuroethical issues in cognitive enhancement. *Journal of Psychopharmacology*, 25(2) :197–204.

Sakai, M. (2015). Modeling the relationship between heart rate and features of vocal frequency. *International Journal of Computer Applications*, 120(6).

Sara, S. J. (2000). Retrieval and reconsolidation : Toward a neurobiology of remembering. *Learning & Memory*, 7(2) :73–84.

Scherer, K. R. (2003). Vocal communication of emotion : A review of research paradigms. *Speech communication*, 40(1) :227–256.

Scherer, K. R., Banse, R., and Wallbott, H. G. (2001). Emotion inferences from vocal expression correlate across languages and cultures. *Journal of Cross-cultural psychology*, 32(1) :76–92.

Scherer, S., Lucas, G. M., Gratch, J., Rizzo, A. S., and Morency, L.-P. (2015). Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews. *IEEE Transactions on Affective Computing*, 7(1) :59–73.

Scherer, S., Stratou, G., Mahmoud, M., Boberg, J., Gratch, J., Rizzo, A., and Morency, L.-P. (2013). Automatic behavior descriptors for psychological disorder analysis. page 1–8, Shanghai, China. IEEE.

Schuller, B., Batliner, A., Seppi, D., Steidl, S., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., and Aharonson, V. (2007). *The relevance of feature type for the automatic classification of emotional user states : low level descriptors and functionals*.

Schuller, B. W., Friedmann, F., and Eyben, F. (2014). The munich biovoice corpus : Effects of physical exercising, heart rate, and skin conductance on human speech production. In *LREC*, pages 1506–1510.

Schwartz, L., Levy, J., Endevelt-Shapira, Y., Djalovski, A., Hayut, O., Dumas, G., and Feldman, R. (2022). Technologically-assisted communication attenuates inter-brain synchrony. *NeuroImage*, 264 :119677.

Searle, J. (2010). *Making the social world : The structure of human civilization*. Oxford University Press.

Seguin, L. and Tassy, L. (2022). E-santé, digitalisation ou transformation numérique : impact sur les soins de support en oncologie. *Bulletin du Cancer*, 109(5) :598–611.

Sherrington, C. (1946). The integrative action of the nervous system. *Cambridge University Press*.

Simondon, G. (1989). *Du mode d'existence des objets techniques*. Aubier.

Simonyan, K. and Horwitz, B. (2011). Laryngeal motor cortex and control of speech in humans. *The Neuroscientist*, 17(2) :197–208.

Stewart, A. M., Lewis, G. F., Heilman, K. J., Davila, M. I., Coleman, D. D., Aylward, S. A., and Porges, S. W. (2013). The covariation of acoustic features of infant cries and autonomic state. *Physiology & Behavior*, 120 :203–210.

Stiegler, B. (2016a). *Dans la disruption. Comment de pas devenir fou ?* Les Liens qui Libèrent.

Stiegler, B. (2016b). *Dans la disruption : Comment ne pas devenir fou ?* Éditions les liens qui libèrent.

Stuart, A., Kalinowski, J., Rastatter, M. P., and Lynch, K. (2002). Effect of delayed auditory feedback on normal speakers at two speech rates. *The Journal of the Acoustical Society of America*, 111(5) :2237–2241.

Stylianou, Y. (2009). Voice transformation : a survey. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 3585–3588. IEEE.

Tanaka, Y., Terasawa, Y., and Umeda, S. (2021). Effects of interoceptive accuracy in autonomic responses to external stimuli based on cardiac rhythm. *PLOS ONE*, 16(8) :e0256914.

Taschereau-Dumouchel, V., Michel, M., Lau, H., Hofmann, S. G., and LeDoux, J. E. (2022). Putting the "mental" back in "mental disorders" : a perspective from research on fear and anxiety. *Molecular Psychiatry*, 27(3) :1322–1330.

Thayer, J. F., Åhs, F., Fredrikson, M., Sollers III, J. J., and Wager, T. D. (2012). A meta-analysis of heart rate variability and neuroimaging studies : implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews*, 36(2) :747–756.

Theriault, J. E., Young, L., and Barrett, L. F. (2021). The sense of should : A biologically-based framework for modeling social pressure. *Physics of Life Reviews*, 36 :100–136.

Thoret, E., Andrillon, T., Gauriau, C., Leger, D., and Pressnitzer, D. (2022). Sleep deprivation measured by voice analysis. *bioRxiv*, pages 2022–11.

Titze, I. R. and Martin, D. W. (1998a). Principles of voice production. *The Journal of the Acoustical Society of America*, 104(3) :1148–1148.

Titze, I. R. and Martin, D. W. (1998b). Principles of voice production.

Toda, T., Black, A. W., and Tokuda, K. (2007). Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory. *IEEE Transactions on Audio, Speech, and Language Processing*, 15(8) :2222–2235.

Trouvain, J. and Truong, K. P. (2015). Prosodic characteristics of read speech before and after treadmill running. In *Sixteenth annual conference of the international speech communication association*.

Tsakiris, M., Jiménez, A. T., and Costantini, M. (2011). Just a heartbeat away from one's body : interoceptive sensitivity predicts malleability of body-

representations. *Proceedings of the Royal Society B : Biological Sciences*, 278(1717) :2470–2476.

Tucciarelli, R., Vehar, N., Chandaria, S., and Tsakiris, M. (2022). On the realness of people who do not exist : The social processing of artificial faces. *iScience*, 25(12) :105441.

Umetani, K., Singer, D. H., McCraty, R., and Atkinson, M. (1998). Twenty-four hour time domain heart rate variability and heart rate : relations to age and gender over nine decades. *Journal of the American College of Cardiology*, 31(3) :593–601.

Usman, M., Zubair, M., Ahmad, Z., Zaidi, M., Ijyas, T., Parayangat, M., Wajid, M., Shiblee, M., and Ali, S. J. (2021). Heart rate detection and classification from speech spectral features using machine learning. *Archives of Acoustics*, 46.

Vaccari, C. and Chadwick, A. (2020). Deepfakes and disinformation : Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society*, 6(1) :2056305120903408.

Vaiva, G., Jehel, L., Cottencin, O., Ducrocq, F., Duchet, C., Omnes, C., Genest, P., Rouillon, F., and Roelandt, J.-L. (2008). Prévalence des troubles psychotraumatiques en france métropolitaine. *L'encéphale*, 34(6) :577–583.

van den Broek, E. L., van der Sluis, F., and Dijkstra, T. (2011). Telling the story and re-living the past : How speech analysis can reveal emotions in post-traumatic stress disorder (ptsd) patients. *Sensing Emotions : The impact of context on experience measurements*, pages 153–180.

van der Burg, S. (2009). Taking the "soft impacts" of technology into account : broadening the discourse in research practice. *Social Epistemology*, 23(3-4) :301–316.

van Gent, P., Farah, H., Nes, N., and Arem, B. (2019). Heartpy : A novel heart rate algorithm for the analysis of noisy signals. *Transportation Research Part F : Traffic Psychology and Behaviour*, 66 :368–378.

Vincent, J. (2019). This ai-generated joe rogan fake has to be heard to be believed. the most realistic ai voice clone we've heard. *The Verge*.

Wadhwa, H., Gradinaru, C., Gates, G. J., Badr, M. S., and Mateika, J. H. (2008). Impact of intermittent hypoxia on long-term facilitation of minute ventilation and heart rate variability in men and women : do sex differences exist ? *Journal of applied physiology*, 104(6) :1625–1633.

Ward, B., Ward, M., Fried, O., and Paskhover, B. (2018). Nasal distortion in short-distance photographs : the selfie effect. *JAMA facial plastic surgery*.

Weathers, F. W., Litz, B. T., Herman, D. S., Huska, J. A., Keane, T. M., et al. (1993). The ptsd checklist (pcl) : Reliability, validity, and diagnostic utility. In *annual convention of the international society for traumatic stress studies, San Antonio, TX*, volume 462. San Antonio, TX;.

Weber, M. and Colliot-Thélène, C. (2003). Le savant et le politique : une nouvelle traduction. *(No Title)*.

Weiss, A., Burgmer, P., and Mussweiler, T. (2018). Two-faced morality : Distrust promotes divergent moral standards for the self versus others. *Personality and Social Psychology Bulletin*, 44(12) :1712–1724.

Westenberg, H. and Sandner, C. (2006). Tolerability and safety of fluvoxamine and other antidepressants. *International Journal of Clinical Practice*, 60(4) :482–491.

Wu, Y.-C., Hayashi, T., Tobing, P. L., Kobayashi, K., and Toda, T. (2019). Quasi-periodic wavenet vocoder : A pitch dependent dilated convolution model for parametric speech generation. *arXiv preprint arXiv :1907.00797*.

Xu, R., Mei, G., Zhang, G., Gao, P., Judkins, T., Cannizzaro, M., Li, J., et al. (2012). A voice-based automated system for ptsd screening and monitoring.

Yang, Y., Fairbairn, C., and Cohn, J. F. (2012). Detecting depression severity from vocal prosody. *IEEE transactions on affective computing*, 4(2) :142–150.

Yoshida, S., Tanikawa, T., Sakurai, S., Hirose, M., and Narumi, T. (2013). Manipulation of an emotional experience by real-time deformed facial feedback. In *Proceedings of the 4th Augmented Human International Conference*, pages 35–42.

Zheng, Z. Z., MacDonald, E. N., Munhall, K. G., and Johnsrude, I. S. (2011). Perceiving a stranger's voice as being one's own : A 'rubber voice'illusion ? *PloS one*, 6(4) :e18655.