Perception de l'aspect naturel de phonèmes produits avec différentes méthodes de synthèse de la parole

Auteur : Fontaine, Camille
Promoteur(s) : Remacle, Angélique; 16572
Faculté : þÿ F a c u l t é d e P s y c h o l o g i e , L o g o p é d i e e t S c i e n c e s d e l E d u c a t i o n
Diplôme : Master en logopédie, à finalité spécialisée en voix
Année académique : 2021-2022
URI/URL : http://hdl.handle.net/2268.2/14310

LIÈGE université
Psychologie, Logopédie
& Sciences de l'Éducation

Département de logopédie

Perception de l'aspect naturel de phonèmes produits avec différentes méthodes de synthèse de la parole

Mémoire présenté en vue de l'obtention du grade de

Master en Logopédie, à finalité spécialisée en voix

Promotrice : Angélique Remacle

Co-promoteur : Rémi Blandin

Camille FONTAINE

Année académique 2021-2022

REMERCIEMENTS

Mes premiers remerciements vont à mes promoteurs, Madame Angélique Remacle et Monsieur Rémi Blandin. Merci pour votre bienveillance, votre disponibilité, votre encadrement, vos encouragements, vos précieux conseils, et votre soutien durant ces deux années.

Je remercie chaleureusement Madame Morgane Warnier et Monsieur Vincent Didone, lecteurs de ce mémoire. Merci pour votre aide et pour le partage de vos connaissances pour la réalisation des analyses statistiques. Merci pour le temps et l'intérêt que vous consacrerez à ce mémoire.

Je remercie également Monsieur Xavier Kaiser. Merci pour votre disponibilité, pour vos réponses à mes questions. Merci d'avoir mis la cabine audiométrique du CEDIA à ma disposition afin de réaliser mes testings dans un environnement favorable.

J'adresse également mes remerciements à l'ensemble des participants pour les pré-tests et les testings. Merci pour votre disponibilité, votre entraide étudiante, votre gentillesse et votre dévouement. Sans vous, ce mémoire n'aurait pas pu voir le jour.

Je remercie aussi mes « logopotes ». Merci pour votre soutien durant ces 5 années d'études. Merci pour les rires, merci pour les souvenirs. Merci d'avoir rendu ce parcours universitaire si mémorable.

Mes derniers remerciements vont vers ma famille. Merci à vous tous. Je remercie mes parents, mes grands-parents et mon frère, Thomas, pour vos encouragements et votre soutien sans faille durant ces 5 années. Merci pour vos conseils, vos aides, vos remontes moral. Vous avez toujours cru en moi, et sans vous, je n'aurais pas pu réaliser ces dernières années. Merci à vous 5 pour votre amour. Un clin d'œil particulier à mon grand-père, mon meilleur ami durant toutes mes études. À nos rires, à nos trajets, à nos examens, à nos sessions, à notre étude. Merci à vous tous pour tout ce que vous avez fait pour moi. Je ne vous remercierai jamais assez.

TABLE DES MATIÈRES

LISTE DES ABRÉVIATIONS

BWE	Algorithmes d'extensions
CSV	Comma separated values
CV	Consonne-voyelle
dB	Décibel
FPLSE	Faculté de psychologie, logopédie et sciences de l'éducation
f_0	Fréquence fondamentale
HD	Haute définition
HF	Hautes fréquences
HFE	Hautes fréquences étendues
Hz	Hertz
IRM	Imagerie par résonnance magnétique
kHz	Kilohertz
MM	Modèle tridimensionnel
NB	Narrow band
F1	Premier formant
F2	Second formant
SPL	Sound pressure level
3D	Tridimensionnel
1D	Unidimensionnel
WB	Wide band
X^2	Khi-Carré

LISTE DES FIGURES

LISTE DES TABLEAUX

I. INTRODUCTION GÉNÉRALE

La parole de synthèse (ou parole artificielle) peut être réalisée avec différentes méthodes comme la synthèse concaténative, la synthèse par apprentissage profond, la synthèse par règles, ou encore la synthèse articulatoire. Cependant, aucune méthode de synthèse ne fournit une parole totalement naturelle (Gully, 2017). Il existe différentes modélisations physiques, notamment le modèle acoustique unidimensionnel (1D), le modèle acoustique tridimensionnel (3D) et le modèle d'algorithme d'extension (BWE). Le modèle 3D tente d'offrir une parole de synthèse se rapprochant le plus de la parole naturelle (Gully, 2017). Il prend en compte la forme précise du tractus vocal ce qui génère des simulations acoustiques proches de la réalité (Arnela et al., 2019 ; Freixes et al., 2018). La synthèse articulatoire du modèle 3D semble encourageante pour étudier les éventuelles contributions des hautes fréquences (HF) dans l'aspect naturel du signal de parole (Arnela et al., 2019 ; Freixes et al., 2018 ; Monson, Hunter et al., 2014). La synthèse articulatoire permet de générer des phonèmes via le contrôle des paramètres des articulateurs (ouverture labiale, position de la hauteur de la langue, ...). Elle a l'avantage de ne pas être influencée par des effets liés au locuteur (Birkholz et al., 2017 ; Delvaux & Pillot-Loiseau, 2019), et est donc idéale pour des expériences de perception auditive. Gully (2017) a évalué les différences perceptives produites par différents modèles acoustiques auprès de juges réels. Cette étude révèle que la parole de synthèse produite avec le modèle 3D est perçue comme plus naturelle (Gully, 2017). Dans ce mémoire, nous avons souhaité investiguer les différences perceptives produites par le modèle 1D, le modèle 3D et le modèle BWE auprès d'auditeurs réels. Nous avons cherché à déterminer si l'un de ces modèles, notamment le modèle 3D, produisait une parole de synthèse plus naturelle que les autres.

Longtemps ignorées dans la majorité des recherches concernant la perception de la parole, les HF (> 5kHz) ont connu un nouvel intérêt cette dernière décennie (Vitela et al., 2015 ; Monson et Caravello, 2019 ; Boyd-Pratt et Donnai, 2020). Plusieurs études ont démontré leur importance, notamment au niveau perceptif (Monson, 2014 ; Monson, 2019a). Les HF semblent jouer un rôle important dans le signal de la parole (Monson et al., 2011 ; Monson & Caravello, 2019). Récemment, des auteurs ont évoqué le potentiel rôle de celles-ci pour offrir un aspect plus naturel au signal de parole (Birkholz & Drechsel, 2021). Dans ce travail, nous avons souhaité approfondir cette question du rôle des HF dans l'aspect naturel de la parole au travers de la parole de synthèse.

Ce mémoire s'inscrit dans le cadre d'un projet de recherche mené par Angélique Remacle (Unité logopédie de la voix, ULiège), Rémi Blandin & ses collaborateurs à l'Université de Dresde (Institute of Acoustics and Speech Communication, Technische Universität Dresden). Ces derniers souhaitent développer un outil de synthèse articulatoire à large bande dont l'aspect serait le plus naturel possible. L'objectif de ce mémoire est de déterminer si l'aspect naturel de la parole produite avec la synthèse articulatoire varie selon la méthode utilisée. Étudier le rôle des HF dans la perception de l'aspect naturel de parole de synthèse implique différents enjeux, notamment, déterminer si les HF jouent réellement un rôle dans la parole de synthèse. Si tel est le cas, l'insertion des HF permettraient de générer des modèles acoustiques davantage complexes pour produire une parole de synthèse. Si ce n'est pas le cas, les modèles acoustiques basiques pour générer de la parole de synthèse seraient suffisants.

Pour répondre à notre objectif, 40 participants ont entendu, lors de deux expériences différentes, un total de 20 stimuli synthétiques : 5 phonèmes ([a], [e], [i], [o], [u] x 2 genres (homme, femme) x 2 types de voix (modale, pressée). Les phonèmes qui seront utilisés dans notre expérience ont été créés avec le logiciel VocalTractLab. Il s'agit d'un logiciel proposant un synthétiseur articulatoire de la parole qui permet de produire des énoncés de haute qualité acoustique (Birkholz, 2013). Chaque stimulus a été entendu au travers de 3 méthodes (1D, 3D et BWE). La première tâche est une comparaison par paires, où les participants sont amenés à comparer deux stimuli (même phonème, même genre, même type de voix, mais modèles acoustiques de synthèses différents) et à décider quel stimulus leur semble le plus naturel. La seconde tâche est une évaluation de l'aspect naturel de chaque stimulus, entendu un à un, sur une échelle métrique allant de 0 à 100. Sur base de la littérature scientifique, les stimuli produits avec le modèle 3D devraient être perçus comme plus naturels grâce à son réalisme acoustique. Notre étude se base sur l'hypothèse selon laquelle les HF sont susceptibles de jouer un rôle dans la perception de la parole et d'améliorer la qualité de la synthèse articulatoire en termes d'aspect naturel.

Nous présenterons plusieurs sections. La première section correspond à l'introduction générale. La seconde consiste en une revue de la littérature où nous aborderons la perception de la parole, les HF et les différentes méthodes de synthèses de parole, toutes trois représentent les thématiques principales de ce mémoire. La troisième évoquera les objectifs et hypothèses. La quatrième section décrira la méthodologie utilisée dans le cadre des deux taches expérimentales. La cinquième section révélera les résultats des expérimentations sur base d'analyses statistiques. Ces résultats seront ensuite discutés et interprétés dans la sixième section.

II. REVUE DE LA LITTÉRATURE

Chapitre 1 - La perception de la parole

1. Distinction des différents concepts

La **parole** découle de la manipulation de la pression d'air générée par le système respiratoire. Elle repose sur un enchaînement rapide de mouvements réalisés par le système respiratoire, les organes de la phonation et les articulateurs qui permettent la coarticulation de différents phonèmes (Baken & Orlikoff, 2000 ; Castellengo, 2015). La **perception de la parole** est le processus par lequel les humains sont capables d'interpréter et de comprendre les sons utilisés dans le langage. La perception de la parole correspond à la capacité d'extraction des unités linguistiques élémentaires (phonèmes) et de leurs caractéristiques acoustiques distinctives dans le signal de la parole (Dieh et al., 2004). Dans ce mémoire, nous nous intéresserons particulièrement à l'aspect naturel de la parole crée via la synthèse articulatoire.

2. Corrélats anatomiques

2.1. Évolution anatomique

Afin de percevoir la parole, divers éléments anatomiques sont indispensables. Certains de ces éléments sont apparus au cours de l'évolution de la lignée humaine. La modification la plus importante est celle qui a eu lieu chez nos ancêtres au niveau de la cochlée. Celle-ci s'est développée davantage tout en abaissant sa limite de fréquence supérieure[1] (Manley, 2016). Cela montre que l'oreille humaine a subi une évolution au fur et à mesure du temps. Cette évolution a été fortement influencée par l'apparition et le développement de la parole (Manley, 2016). La cochlée humaine permet d'entendre des sons entre 20 Hz et 20 kHz (Rebillard, 2021). Celle-ci permet donc d'entendre les sons de la parole, mais aussi les HF. Nous parlerons des HF dans le chapitre II de la présente partie.

1 La comparaison des données humaines et des données de primates montre que la sélectivité humaine est plus élevée aux fréquences inférieures à environ 2 kHz, mais progressivement plus basse à des fréquences de plus en plus élevées. (Manley, 2016)

2.2. Caractéristiques acoustiques essentielles

2.2.1. Perception de la parole

L'oreille humaine peut entendre jusqu'à 20 kHz, ce qui est non-négligeable pour la perception de la parole. En effet, cela permet de distinguer des composants sonores dont la fréquence varie, par exemple, les formants. Ceux-ci sont des bandes de fréquences dans lesquelles se concentre l'énergie acoustique. Les voyelles sont caractérisées par la présence de zones d'harmoniques renforcées appelées « formants » (Meunier, 2007). Les 2 premiers formants permettent de discriminer les voyelles. Le premier formant (= F1) est défini selon « le degré d'aperture de la mandibule » (Meunier, 2007, p. 3) et le second formant (= F2) est déterminé selon la position de la langue (avant ou arrière) et selon la position labiale (lèvres étirées, lèvres arrondies, …) (Meunier, 2007). Dans le tableau ci-dessous (Meunier, 2007), nous retrouvons les valeurs formantiques moyennes des voyelles orales du français.

		F1	F2
voy. fermées	i	308	2064
	y	300	1750
	u	315	764
voy. mi-fermées	e	365	1961
	ø	381	1417
	o	383	793
voy. mi-ouvertes	ɛ	530	1718
	œ	517	1391
	ɔ	531	998
voy.ouv.	a	684	1256

Tableau 1 : Valeurs formantiques des voyelles orales (Meunier, 2007)

En réalité, discriminer et reconnaitre la parole est l'une des fonctions les plus élémentaires du système auditif humain, notamment dans un contexte où l'on cherche à identifier un locuteur (Belin et al., 2004 ; Van Dommelen, 1990). Cette discrimination est possible grâce aux caractéristiques acoustiques et linguistiques du locuteur (Baumann, 2008). Ces caractéristiques sont attribuables aux différences anatomiques des structures vocales et aux différences d'utilisation du système vocal (Bricker & Pruzansky, 1976 ; Hecker, 1971). De plus, la reconnaissance et la différenciation des voix sont réalisables grâce aux caractéristiques prosodiques, c'est-à-dire, grâce à la hauteur, l'intensité, et la durée (Baumann, 2008). Des mesures acoustiques telles que

les fréquences des formants et la fréquence fondamentale[2] (f_0) jouent également un rôle dans la perception de la parole (Baumann, 2008). En 1978, Walden et ses collaborateurs ont présenté un modèle perceptif à 4 dimensions (Walden et al., 1978). Ce dernier mettait en avant la f_0, l'âge du locuteur, la durée du mot, et la qualité vocale. En 1978, d'autres auteurs ont également précisé que le genre du locuteur était une variable à prendre en compte dans la discrimination de la parole (Singh & Murry, 1978). Baumann (2008) confirme cette information, en ajoutant que les différences au niveau de la glotte[3] et des plis vocaux liées au genre influencent également la perception de la parole. En outre, des analyses statistiques antérieures ont confirmé que pour discriminer la parole féminine par exemple, il fallait se fier à la différence entre les fréquences des formants et les caractéristiques glottiques (Hanson, 1997). Un effet lié au genre du locuteur est donc constaté. Nous y reviendrons plus tard dans ce travail (II, chap. 2, 4.2).

2.2.2. Types de voix

Il existe plusieurs types de voix (Gordon, 2002) : la voix modale, la voix soufflée et la voix pressée, que nous allons détailler ci-dessous. Ces différents types de voix s'inscrivent sur un continuum allant de « la voix soufflée, à la voix modale, à la voix pressée » (Titze, 2000, p. 248) en fonction du degré d'adduction glottique. Nous allons à présent définir les trois types de voix, bien que dans ce mémoire, seule les voix modales et les voix pressées ont été utilisées dans notre expérimentation. Nous y reviendrons dans notre 4ème partie.

2.2.2.1.Voix modale

La voix modale correspond « à une vibration régulière des plis vocaux tout au long de leur surface » (Ioannidis et al., 2014, p. 3). Cette vibration a une grande amplitude globale, il en résulte une voix avec un timbre relativement plus riche (Titze, 2000). Il s'agit de la voix la plus fréquente. La voix modale se caractérise par une « tension adductive modérée » (Wright et al., 2019, p. 4) et par un quotient ouvert[4] modéré (Wright et al., 2019). La résistance laryngée est modérée (Wright et al., 2019).

2 La fréquence fondamentale est l'harmonique de premier rang d'un son.

3 La glotte correspond à l'espace entre les plis vocaux. (Baumann, 2008)

4 Le quotient ouvert est « un paramètre lié à la source glottique. Il est défini par le rapport de la durée d'ouverture par la période. Il varie donc de 0 (cas théorique d'un cycle glottique sans ouverture) à 1 (cas d'une fermeture incomplète) » (Lamesch, 2006, p. 24).

2.2.2.2.Voix soufflée

La voix est dite soufflée lorsque, durant la vibration, les plis vocaux sont partiellement ouverts et lorsque l'adduction de ceux-ci est maitrisée par des muscles détendus, permettant à l'air de se répandre à travers la glotte (Sundberg, 1987). La voix soufflée correspond à une « hypo-adduction » (Titze, 2000, p. 275). Cette voix résulte d'une « faible pression sous-glottique combinée à un débit glottique élevé, excessif » (Titze, 2000, p. 84). Les plis vocaux ne sont que faiblement en contact, tandis que le quotient d'ouverture est grand, ce qui implique une faible résistance laryngée (Lei et al., 2019) Le rendu de ce type de voix ressemble à un soupir, un chuchotement.

2.2.2.3.Voix pressée

Une voix pressée ou serrée est produite quand une tension forte est appliquée à la musculature qui entoure le larynx. Ce mode phonatoire induit un excès d'adduction glottique (Remacle, 2013), on parle d'ailleurs d' « hyper-adduction » (Titze, 2000, p. 275). Dans ce cas, l'ensemble du larynx est soumis à une tension et le haut du larynx peut se contracter fortement. Les plis vocaux ont une vibration inefficace et le débit d'air devient plus faible (Sundberg, 1995), voire insuffisant (Titze, 2000). Les plis vocaux sont plus souvent en contact, et le quotient d'ouverture est plus petit, ce qui implique une forte résistance laryngée (Lei et al., 2019). Dans ce cas, seule une petite quantité d'air parvient à passer à travers les plis vocaux. La voix qui en découle est une voix tendue, serrée et restreinte en résonnance.

3. La sensibilité auditive

3.1. La perception des sons

Toutes les fréquences que nous entendons ne sont pas perçues de la même façon et avec la même intensité (Antoine, 2019). La *Figure 1* des lignes isosoniques normales le démontre. En effet, l'analyse de cette figure montre qu'un niveau sonore de 70 dB minimum est nécessaire pour percevoir une fréquence de 20 Hz. Quant aux fréquences situées entre 1000 et 5000 Hz, nous pouvons voir qu'il est plus aisé de les entendre étant donné que le niveau de pression sonore qui y est lié est proche de 0. L'étendue fréquentielle de la parole s'étend essentiellement de 78 Hz à 5000 Hz (Sicard, 1995).

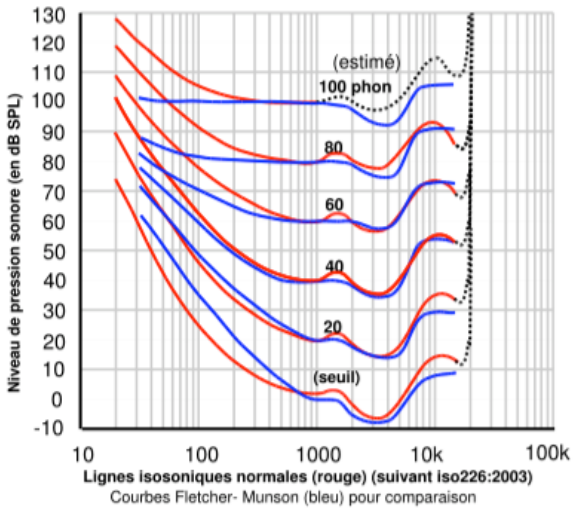

Figure 1 : Lignes isosoniques normales (Vallot, 2019)

3.2. Audiométrie tonale

L'audiométrie tonale est un examen permettant d'évaluer l'acuité auditive en évaluant la conduction aérienne et la conduction osseuse. Il s'agit d'un examen bénin facile à administrer. Toutefois, cette simplicité est trompeuse car elle résulte d'une bonne coopération du participant (Maci & Carusi, 2020). Néanmoins, l'audiométrie tonale classique ne permet pas d'évaluer les HF. Cela a une importance particulière dans ce mémoire, nous développerons ce point dans notre chapitre 2 sur les HF (II, chap. 2, 1).

L'audiométrie tonale se déroule en cabine insonorisée. Différents stimuli sonores sont administrés au patient par le biais d'écouteurs. L'audiométrie doit être effectuée avec un niveau de bruit ambiant le plus faible possible pour mesurer un seuil d'audibilité qui tend à se rapprocher de 0 dB (Maci & Carusi, 2020).

La perte auditive se calcule sur les fréquences 0.5 kHz, 1 kHz, 2 kHz et 4 kHz. Une moyenne des scores obtenus pour chaque fréquence est réalisée. La valeur qui est obtenue suite à ce calcul permet de classifier l'audition. Si la perte est

- de 0 à 20 dB alors l'audition est considérée comme normale ;
- de 21 à 40 dB, nous sommes face à une surdité légère ;
- de 41 à 70 dB, nous sommes face à une surdité moyenne ;
- de 71 à 90 dB, nous sommes face à une surdité sévère ;
- au-delà de 91 dB, nous sommes face à une surdité profonde (Recommandation biap 02/1 bis : Classification audiométrique des déficiences auditives).

3.3. Perte auditive

L'un des points importants pour l'audiologie est la notion « d'audition normale » qui est observable via l'audiométrie tonale pure (Moore, 2017). Il existe beaucoup de facteurs qui affectent l'audition, comme l'âge (Zadeh, 2019), l'exposition au bruit ou encore la prise de certains médicaments (Rodríguez-Valiente et al., 2016).

La distribution spectrale[5] de l'énergie joue un rôle dans la perception de la parole, au travers d'informations indispensables pour le signal de la parole, telles que la suppression d'une bande de parole, la limite supérieure de l'audition humaine (environ 20 kHz) et la sensibilité auditive humaine (Lippman, 1996 ; Best et al., 2005 ; Vitela et al. 2015). Ce point sera mis en lien avec les HF et détaillé dans le chapitre suivant (II, chap. 2, 1).

En résumé, la parole est produite par différents mouvements réalisés par le système respiratoire, les articulateurs et les organes de la phonation. L'oreille humaine a fortement évolué avec l'apparition de la parole, par conséquent, la perception de celle-ci permet aux Hommes de comprendre et d'interpréter le signal de parole. Notons que la perception de la parole dépend de la sensibilité auditive des auditeurs mais aussi de plusieurs caractéristiques acoustiques (fréquence des formants, fréquence fondamentale), anatomiques (genre, structures vocales, utilisation du système vocale), linguistiques (prononciation, accent) et prosodiques (hauteur, intensité, durée) du locuteur.

5 La distribution spectrale est un graphique qui représente l'énergie (la force) de chaque longueur d'ondes.

Chapitre 2 - Les hautes fréquences

1. Définition

Dans le contexte de la parole, les hautes fréquences étendues (HFE) se réfèrent aux fréquences qui s'étendent de 5.7 kHz à 22 kHz (Monson, 2014). Comme nous pouvons le constater sur la *Figure 2*, la nature et la quantité d'énergie des hautes fréquences diffèrent selon qu'il s'agisse d'une consonne ou d'une voyelle. Les consonnes sont plus riches en hautes fréquences que les voyelles.

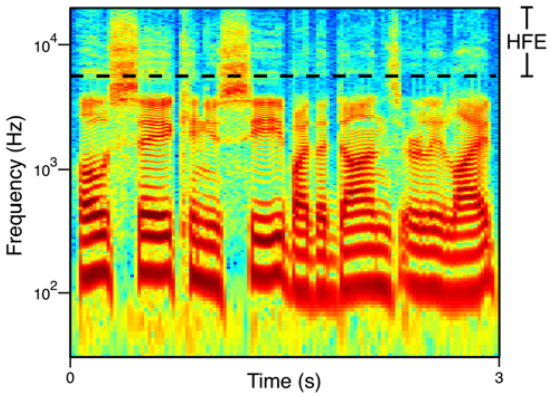

Figure 2 : Spectrogramme de la phrase « Oh say can you see by the dawn's early light » (Monson, Lotto et al., 2014) prononcée par un homme. Sur cette figure la délimitation de la plage des HFE (les bandes d'octaves 8 et 16 kHz) est mise en évidence par une ligne en pointillés.

Les HF correspondent à une région du spectre auditif humain souvent ignorée par les cliniciens et les chercheurs (Hunter, 2020 ; Monson, 2011 ; Vitela, 2015). En effet, nous pouvons entendre jusqu'à 20 kHz (Monson & Caravello, 2019 ; Zadeh, 2019) et pourtant l'audiométrie utilisée en clinique mesure uniquement la sensibilité auditive jusqu'à 8 kHz (Hunter, 2020 ; Moore, 2017 ; Zadeh, 2019). Cependant, il existe, en plus de l'audiométrie tonale conventionnelle (125 Hz à 8000 Hz), une audiométrie pour les HFE allant de 9.000 Hz à 20.000 Hz (Rodríguez-Valiente et al., 2016). Toutefois, cette dernière n'est que rarement utilisée en clinique, bien qu'elle montre l'importance indispensable des HF (Zadeh, 2019). L'audiométrie tonale conventionnelle reste préférée par les cliniciens et les chercheurs car elle permet un gain de temps et ne demande pas un matériel spécifique. En effet, celle-ci permet de ne cibler que les fréquences les plus importantes auxquelles nous sommes le plus sensibles et permet d'éviter les difficultés techniques liées à la mesure des HFE. Néanmoins, Brian Monson (2014) émet l'hypothèse que si l'être humain dispose des HF, c'est pour un but précis. Par exemple pour détecter des proies, des prédateurs, pour percevoir la parole dans des conditions difficiles ou

encore pour discriminer les consonnes (Monson, 2014). Notons que l'énergie des HF de la parole, par exemple, est utilisée pour entendre la parole dans des environnements bruyants (Zadeh, 2019). Cela signifie que parfois, certains audiogrammes sont intacts mais que les auditeurs perçoivent mal la parole dans le bruit (Badri, 2011). En effet, la perception de cette dernière diminue avec l'âge, sans pour autant affecter la qualité d'un audiogramme. Il s'agit d'un élément important pour notre partie méthodologie, nous en reparlerons dans la section qui y est consacrée.

2. Rôles des hautes fréquences

2.1. Aspect naturel de la parole

L'aspect naturel de la parole est un élément essentiel dans ce mémoire, puisque nous cherchons à évaluer l'aspect naturel de la parole avec des stimuli de synthèses. Cependant, déterminer l'aspect naturel de la parole est très compliqué pour les humains (Gully, 2017). Les humains disposent chacun de critères internes pour juger de l'aspect naturel de la parole, il n'est pas possible de mesurer le degré de naturel objectivement. La seule solution est de demander à plusieurs auditeurs humains de juger différents stimuli en les comparant (Gully, 2017). Nous avons pris en compte cela pour la réalisation de notre première expérience, nous y reviendrons (IV, chap. 3, 1.1).

La perception de l'aspect naturel de la parole est affectée par les HF. Moore & Tan (2003) ont étudié cet aspect naturel de la parole à l'aide d'une tâche de « *jugements subjectifs de qualité* » sur des signaux de musique et de parole. Ils ont soumis plusieurs bandes passantes à 10 auditeurs en leur demandant de les comparer et de dire si elles étaient naturelles ou non. Une bande avec un filtre passe-bas à 10.9 kHz et une bande passante limitée à 7 kHz ont été comparées par les participants. Selon eux, la bande à 10.9 kHz semblait plus naturelle. Il leur a ensuite été administré deux autres bandes fréquentielles, une limitée à 10.9 kHz et l'autre à 16.9 kHz. Dans ce cas, aucune différence perceptive n'a été relevée par les participants de l'étude. Les résultats de cette étude démontrent que les HF semblent donner un aspect davantage naturel à la parole. Elle souligne donc l'importance des HF dans l'aspect naturel de la parole. En effet, les HF sont présentes dans la parole naturelle et contiennent des informations importantes (Best, 2005).

2.2. Intelligibilité de la parole

L'intelligibilité de la parole est affectée par les HF (Badri et al., 2011; Moore et al., 2010; Pittman, 2008; Apoux et Bacon, 2004 ; Monson, 2014). De récentes recherches avancent l'utilité des HF dans la compréhension de la parole (Monson, 2012a). L'énergie des HF offre une signification perceptive (Monson, 2011 ; Zadeh, 2019) car elle affecte la perception de l'intelligibilité (Monson, 2012a ; Vitela, 2015 ; Zadeh, 2019). Les résultats de Vitela et de ses collaborateurs (2015) démontrent que les HF offrent des informations sur la distinction des consonnes-voyelles, sur le lieu d'articulation, ainsi que sur la façon de percevoir les sons et de les discriminer. Sans les HF, certaines consonnes seraient inintelligibles, et ce, de façon encore plus prononcée dans des environnements bruyants (Pulakka, 2012 ; Moore & Tan, 2003). Par conséquent, elles jouent un rôle important dans l'intelligibilité de la parole. Les HF sont présentes à différents niveaux dans la parole de tous les jours (Monson, 2014). C'est grâce à elles que nous pouvons percevoir et comprendre certaines informations acoustiques.

2.3. Localisation et directivité de la parole

La propagation des HF se marque de façon directionnelle (Vitela, 2015). En effet, plus la fréquence augmente, plus l'énergie acoustique devient directionnelle (Monson, Lotto, et al., 2014). La directivité caractérise la capacité d'un émetteur ou d'un récepteur à exercer sa fonction dans les différentes directions.

Certains neurones corticaux réglés sur des HF (supérieures à 6 kHz) répondent spécifiquement lors des tâches vocales concurrentes (Mesgarani & Chang, 2012 ; Monson, 2014). Des recherches montrent que les HF jouent un rôle dans la localisation de la parole (Monson et al., 2012b ; Mesgarani & Chang, 2012). En outre, les HF améliorent la perception de la parole dans des conditions extrêmes, comme dans du bruit ou des pièces très réverbérantes (Vitela, 2015).

Comme illustré sur la *Figure 3*, les diagrammes de rayonnement des HF sont hautement directionnels. Cela signifie que l'énergie des HF est plus importante en face d'un locuteur que sur les côtés et les derrières (Hunter, 2020). Cela est observé par Monson et al. (2012b) qui met en évidence la richesse des informations que peuvent nous offrir les HF. Cette information sera

importante pour notre partie méthodologie et la façon dont nous procéderons à nos expérimentations.

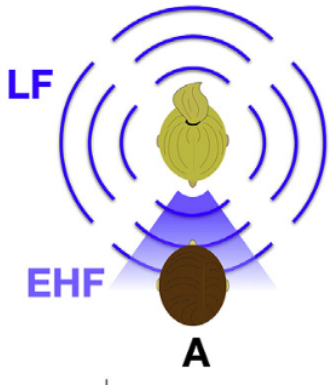

Figure 3 : Propagation des fréquences émises par la parole (Hunter, 2020)

Cette figure montre que les basses fréquences (LF) se dissipent dans toutes les directions autour du locuteur (en vert) alors que les hautes fréquences (HFE), quant à elles, se propagent principalement vers l'avant. Le Point A, quant à lui, représente la position de l'auditeur.

L'audition des HF permet une meilleure discrimination de la rotation de la tête d'un locuteur (Hunter, 2020 ; Monson et al., 2019). En d'autres termes, grâce aux HF, il est possible de localiser la source sonore. La discrimination de l'orientation de la tête du locuteur a été démontrée avec des tâches d'identification de l'orientation de la tête grâce à un haut-parleur rotatif (Imbery et al., 2019). Ainsi, l'énergie des HF sert d'indicateur indispensable afin de définir la direction dans laquelle un locuteur est positionné. Cette capacité est importante pour déterminer si l'on est le destinataire prévu d'un énoncé (Neuhoff, 2003). Cela signifie que les HF ont un effet positif sur la perception spatiale et la localisation de la parole. Elles jouent un rôle indispensable dans des situations où se trouvent beaucoup de sources concurrentes (Best, 2005 ; Rodríguez-Valiente et al., 2014).

2.4. Base pour le développement du langage

Les HF permettent une meilleure compréhension des différents signaux acoustiques dans la parole. Cela est démontré dans l'apprentissage du langage. Des études suggèrent que les enfants utilisent les HF durant leur développement langagier (Pittman, 2008). Les HF facilitent l'apprentissage de la parole (Berlin, 1982). En effet, elles sont essentielles à l'apprentissage car elles permettent la discrimination entre les consonnes qui se distinguent par l'énergie HFE (Monson, 2014 ; Hunter, 2020). En outre, il a été démontré que des enfants possédant une

sensibilité supérieure dans la gamme HFE, pouvaient apprendre des mots de façon accélérée (Pittman, 2008; Stelmachowicz et al., 2007). La qualité du développement du langage repose sur les HF. Lorsque les HF ne sont pas présentes, la qualité du développement langagier est moindre (Moore et Tan, 2003).

Les nourrissons sont extrêmement sensibles aux HF ce qui améliore leur perception des phonèmes et leur offre un meilleur apprentissage du langage (Stelmachowicz et al., 2007). En effet, cette sensibilité est cruciale pour permettre une bonne discrimination des différents sons de la parole, se distinguant par l'énergie des HF (Pittman, 2008).

3. Historique

La mise de côté des HF dans le passé peut s'expliquer, d'une part, par le manque de moyens technologiques permettant de se focaliser sur les HF. D'autre part, il existait un intérêt prédominant pour les basses fréquences (< 5 kHz) qui semblaient être suffisantes pour la perception de la parole (Monson, 2014 ; Monson & Caravello, 2019). Ces deux causes sont à l'origine d'un manque d'informations sur les HF qui se fait actuellement ressentir au sein de la communauté scientifique. Nous allons donc, au sein de cette partie, explorer ces causes plus en profondeur.

3.1. Manque d'informations sur les hautes fréquences

Auparavant, les chercheurs en sciences de la parole et de l'audition ne possédaient que peu voire pas de connaissances sur les HF. En effet, ils pensaient que l'énergie des basses fréquences (de 500 Hz à 4000 Hz) était suffisante pour reproduire une parole de synthèse intelligible (Fletcher et Steinberg, 1930 ; Fletcher et Galt, 1950 ; Rodríguez-Valiente et al., 2016). Même si certains chercheurs s'y sont intéressés dans la première moitié du 20ème siècle, la majorité des études de l'époque ont laissé de côté tout ce qui concernait les HF, les jugeant sans utilité. Cela a contribué au manque de connaissance en ce qui concerne les structures acoustiques dans le signal vocal des HF (Monson, 2019a).

Certaines raisons relevant des facteurs acoustiques peuvent également expliquer le manque de recherches sur les HF. En effet, une inspection du spectre de parole typique révèle que l'énergie

acoustique à des fréquences supérieures à 5 kHz est plus faible (de 20 à 40 dB vers le bas) que celles des basses fréquences du spectre (Moore et al., 2008; Monson et al., 2012a).

3.2. Manque de connaissance sur la technologie

D'autres raisons, quant à elles, remontent au début et au milieu du 20ème siècle. C'est à cette époque que la plupart des travaux de base de la recherche actuelle en sciences de la parole ont été réalisés (Monson, 2014). En effet, l'apparition du téléphone en 1876 et la généralisation de la technologie téléphonique durant le 20ème siècle ont apporté énormément de sujets de recherches. Un des grands intérêts était d'améliorer l'intelligibilité des conversations. Le but était d'améliorer l'efficacité du téléphone sans empiéter sur l'intelligibilité. Cela a été réalisé par le biais des tests d'articulation. En effet, à l'époque des guerres mondiales, il était nécessaire d'obtenir des communications intelligibles dans des environnements acoustiques extrêmement pauvres. Il fallait par conséquent déterminer quels aspects de la parole étaient suffisants et nécessaires pour une conversation intelligible (Monson, 2014). À cette époque, de nombreuses recherches ont eu lieu afin d'étudier la distribution de l'énergie acoustique de la parole et la dépendance en fréquence de l'intelligibilité de la parole au moyen de l'indice d'articulation[6] (Fletcher et Steinberg, 1930 ; Fletcher et Galt, 1950). Cet indice d'articulation, aujourd'hui appelé indice d'intelligibilité, n'attribue que peu de poids aux HF (Monson & Caravello, 2019). Les résultats de ces différentes études ont démontré que les basses fréquences (< 4 kHz) étaient suffisantes pour l'intelligibilité de la parole. Cela est soutenu par d'autres chercheurs qui confirment qu'une énergie inférieure à 7 kHz est suffisante pour reproduire une parole intelligible pour la transmission des systèmes de communication (Fletcher et Galt, 1950; Fletcher et Steinberg, 1930 ; Monson et al., 2014). Ces résultats ont également prouvé que les basses fréquences suffisaient pour étudier l'intelligibilité de la parole. Il n'était donc pas nécessaire d'étudier les HF. En outre, cela concordait parfaitement avec le manque d'avancée technologique qui ne permettait pas d'étudier les HF. En 1922, Crandall et MacKenzie expliquent l'impossibilité de mesurer les HF avec les appareils utilisés à cette époque (Monson et al., 2014).

6 L'indice d'articulation est un indice permettant de mesurer l'intelligibilité de la parole. Il s'étend de 0 à 1. Plus il est proche de 0, plus la parole est inintelligible. Alors que lorsque l'on se rapproche de 1, nous avons une excellente intelligibilité. On parle aujourd'hui d'indice d'intelligibilité.

4. Nouvel intérêt

Récemment, des preuves semblent indiquer que les HF jouent un rôle important dans la perception de la parole et de la voix, contrairement à ce qu'il était pensé autrefois (Monson, 2014 ; Monson & Caravello, 2019). À l'heure actuelle, nous sommes dans une ère où la technologie est omniprésente avec de nombreuses avancées technologiques qui permettent de s'intéresser à ces HF. Ces dernières démontrent leur intérêt au niveau perceptif (Monson, 2014 ; Monson, 2019a).

Les HF sont utilisées dans la parole pour de nombreuses applications, par exemple, par les ingénieurs du son qui manipulent ces HF, notamment pour les voix chantées et les voix parlées. Les auditeurs sont sensibles aux HF et aux changements qu'elles engendrent (Monson, 2014). Aujourd'hui, les téléphones ne sont plus limités aux basses fréquences (< 5 kHz). Les technologies téléphoniques utilisent des « larges bandes » ou des « voix HD[7] ». L'utilisation des « larges bandes » dans la téléphonie signifie que désormais, la bande passante est de 7 kHz, et inclut donc les HF (Geiser, 2012). Cela est également intégré dans les applications de communication numérique, permettant de réaliser des visio-conférences, comme *Skype* par exemple (Geiser, 2012 ; Pulakka et al., 2012).

4.1. Utilité écologique

Les HF sont importantes pour l'homme. En effet, elles offrent une utilité écologique pour la perception de la parole, notamment dans des situations d'écoute difficiles par exemple, lors d'un cocktail, d'une réception, d'une fête, d'un évènement sportif (Hunter, 2020). Dans ces situations, il y a un niveau de bruit ambiant important et de nombreuses personnes différentes qui parlent en même temps. Grâce à leur directivité plus prononcée, les HF vont transmettre les informations du locuteur cible de par l'énergie spectrale. Les basses fréquences vont permettre de par leur énergie de masquage, d'ignorer le bruit de fond et les autres personnes qui parlent autour (Hunter, 2020). Grâce aux HF, comme représenté sur la *Figure 4*, il est possible de détecter et d'isoler l'interlocuteur qui nous intéresse par rapport aux autres interlocuteurs

7 Voix hautes définitions. La communication vocale HD est comprise comme la transmission avec une bande passante audio d'au moins 7 kHz c-à-d de 0 à 7 kHz (Geiser, 2012).

générant le bruit de fond. En outre, elles permettent d'améliorer et d'augmenter l'accès aux informations phonétiques (Hunter, 2020).

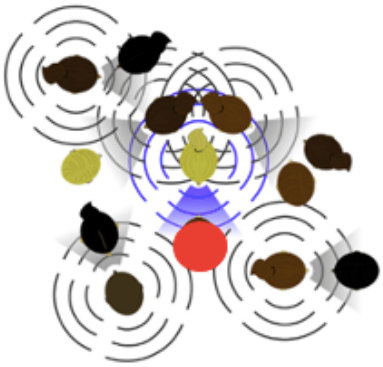

Figure 4 : L'effet Cocktail party (Hunter, 2020)

Cette figure illustre l'effet cocktail party. Les basses fréquences (en arc de cercle) rayonnent dans toutes les directions autour du locuteur (en jaune au centre de la figure). Il y a donc interférences et confusions entre les différents locuteurs. Alors que les hautes fréquences (en bleu), quant à elles, permettent à l'auditeur (en rouge) de détecter l'interlocuteur qui l'intéresse.

Nous allons désormais, par le biais d'un développement dans les points ci-dessous, découvrir que les HF vont influencer la perception et la production de la parole en fonction de plusieurs paramètres.

4.2. Effet du genre du locuteur sur la perception et la production de la parole

Les HF seraient plus importantes pour la perception de la parole des femmes et des enfants que pour celle des hommes (Stelmachowicz et al., 2001). La présence de fréquences aiguës dans la parole suffit à un auditeur pour déterminer le genre du locuteur auquel il est confronté (Monson et al., 2012b).

La f_0 est un paramètre principal pour différencier le genre des locuteurs (Baumann, 2008). Certains chercheurs estiment que les femmes ont tendance à avoir une parole contenant davantage de HF que les hommes (Monson et al., 2012a). Cependant, d'autres chercheurs avancent la variabilité du niveau de HF entre les genres. En effet, la quantité d'énergie des HF est différente entre les paroles féminines et les paroles masculines en fonction de la bande d'octave[8] (Dunn & White, 1940 ; Moore et al., 2008). Dans l'étude de Moore et al., (2008), les

chercheurs ont comparé la parole féminine à la parole masculine sur base de différentes bandes d'octaves[8]. Des enregistrements à large bande passante de la parole conversationnelle normale ont été obtenus à partir d'un échantillon d'hommes et de femmes qui parlaient. Sur base de ces enregistrements, la forme spectrale moyenne sur une large plage de fréquences a été déterminée ainsi que la distribution des niveaux en fonction de la fréquence centrale. Les résultats ont montré que le spectre pour la parole féminine était inférieur à celui de la parole masculine pour les basses fréquences (100 et 125 Hz). Il a également été démontré que pour les fréquences centrales de 200 à 500 Hz, le spectre de la parole féminine était semblable à celui de la voix masculine (étant entre 30 et 41 dB). Lorsque les chercheurs augmentaient à nouveau la fréquence, le spectre de la parole féminine était à nouveau supérieur à celui de la parole masculine. Cela est confirmé par d'autres auteurs qui ont également démontré que le spectre de la parole féminine est supérieur au spectre de la parole masculine, à mesure que la fréquence augmente (Monson et al., 2012a). Par ailleurs, de nombreuses recherches ont prouvé que les voix féminines sont plus riches en HFE que les voix masculines (Dunn & White, 1940 ; Moore et al., 2008 ; Stelmachowicz et al., 2001). Nous pouvons le constater sur la *Figure 5*, ci-dessous. Il y apparait clairement que la voix féminine est, en général, plus riche en HF que celle de l'homme. En effet, nous voyons que les niveaux de pression sonore (SPL) moyens féminins dans des phrases étaient plus élevés que les niveaux masculins pour cinq des six bandes HFE dans la parole et quatre des six bandes HFE dans le chant.

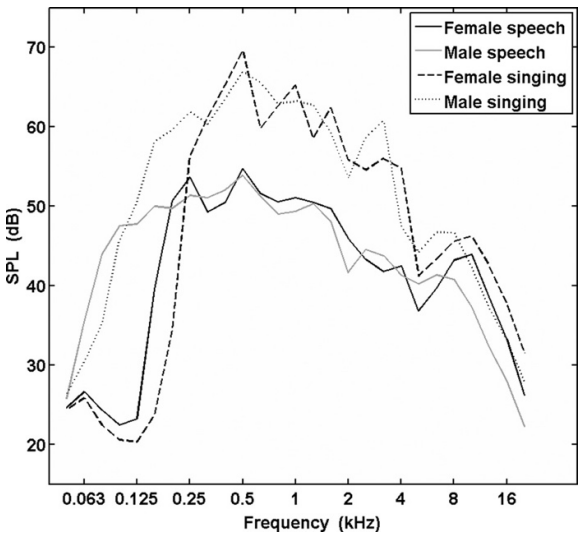

Figure 5 : Graphique des niveaux de pressions sonores féminins et masculins (Monson et al., 2012a)

En abscisse, nous avons la fréquence mesurée en kHz et en ordonnée, la pression sonore (SPL dB).

8 Bande de l'échelle des fréquences permettant l'analyse d'un bruit.

Cette différence entre les paroles masculines et les paroles féminines peut s'expliquer par l'anatomie de l'appareil phonatoire humain. En effet, les hommes et les femmes n'ont pas la même masse de plis vocaux, ni la même longueur de conduit vocal (Smith, 2005). Cela signifie que pour le même son à produire, il existe des différences selon le genre du locuteur, expliquant ainsi que la parole féminine contienne davantage de HF que la parole masculine.

Ce point est important à développer, puisqu'il correspond à l'une des questions de recherches que nous présenterons plus tard (III, chap. 2). Sur base de ces informations, nous souhaitons déterminer si, dans ce mémoire, les stimuli masculins et féminins seront perçus différemment grâce à la génération des HF.

4.3. Effet de l'âge sur la perception des hautes fréquences

L'audition diminue au fur et à mesure avec l'âge. Néanmoins les premiers seuils auditifs à diminuer avec l'âge concernent principalement les HF, suivies progressivement par la diminution des seuils auditifs comprenant des fréquences de plus en plus basses (Rodríguez-Valiente et al., 2014). La perte auditive liée à l'âge dans la perception des HF commence dès le jeune âge adulte (entre 20 et 25 ans), avec des pertes substantielles des HF à 50 ans pour la population générale vieillissante (Green et al., 1987 ; Stelmachowicz et al., 1989).

Les enfants et les adolescents ont une audition particulièrement sensible dans la région des HF, bien que cette sensibilité diminue progressivement au début de l'âge adulte, vers 20 ans (Rodríguez-Valiente et al., 2014 ; Schechter et al., 1986). En effet, déjà chez les jeunes adultes dans la vingtaine, la perception des HF commence à s'atténuer petit à petit, et ainsi diminuer tout au long de la vie (Zadeh, 2019). Rodríguez-Valiente et ses collaborateurs (2014) ont comparé plusieurs groupes de jeunes personnes afin d'observer la différence au niveau des seuils auditifs concernant les HF. Leur résultat met en évidence une meilleure perception des HF pour les participants très jeunes (5 ans à 19 ans) par rapport aux participants étant de jeunes adultes (20 ans à 29 ans). Comme le montre la *Figure 6*, le groupe des 20 à 29 ans présentait des seuils inférieurs à 12.5 kHz et 16 kHz (12 dB et 13.86 dB respectivement) par rapport aux 5-19 ans. Cela confirme la diminution de la sensibilité aux HF au début de la vingtaine. On constate également sur la *Figure 6* que le groupe des 20-29 ans avait besoin de davantage de dB pour percevoir les HF (au-delà de 18 kHz).

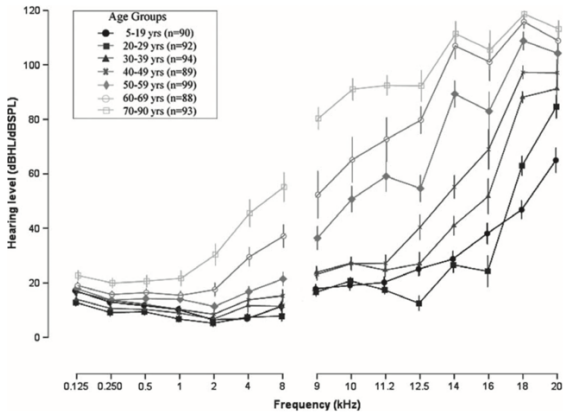

Figure 6 : Seuils d'audition moyens en fonction de la fréquence (kHz) pour différents groupes d'âge. (Rodríguez-Valiente et al., 2014)

Figure représentant les seuils d'audition moyens en fonction de la fréquence (kHz) pour différents groupes d'âge.

Cet effet de l'âge a été démontré dans plusieurs études, et les résultats de celles-ci convergent. Il est vrai que les enfants ont un meilleur accès aux seuils et aux informations de tonalité des HF que les adultes ou les personnes âgées (Green et al., 1987 ; Stelmachowicz et al., 1989).

L'âge est particulièrement important à décrire dans le cadre de ce mémoire. En effet, afin de mettre en place nos expérimentations, nous avons dû établir des critères d'inclusion afin de trouver des participants. Sur base de la littérature évoquée, nous avons décidé que les participants n'auraient pas plus de 25 ans, puisque l'on voit que c'est autour de cette tranche d'âge que la perception des HF commence à diminuer.

4.4. Effet du type de phonème

Les fréquences varient selon les phonèmes. De la sorte, l'énergie des voyelles se situe principalement entre 250 - 2000 Hz. L'énergie des consonnes est quant à elle différente. Les consonnes voisées *([b],[d],[m])* développent une énergie comprise entre 250 - 4000 Hz. Alors que les consonnes non voisées *([f],[s],[t])* se situent dans la gamme de fréquences comprises entre 2000 - 8000 Hz (Alexander, 2014). Il existe également des différences entre les voyelles. Comme nous venons de le voir, la nature et la quantité d'énergie des HF diffèrent selon qu'il s'agisse d'une consonne ou d'une voyelle. Les consonnes sont plus riches en HF que les voyelles (Monson, 2014). Les consonnes, en particulier les fricatives, semblent avoir un niveau de HF plus élevé que les voyelles (Monson, 2011 ; Jongman, 2000). En outre, les consonnes fricatives sourdes semblent être la classe de consonnes qui domine le plus le spectre de la parole

en HF (Monson et al., 2012a ; Rodríguez-Valiente et al., 2016) par rapport aux fricatives sonores qui sont un peu moins proéminentes (Jongman et al., 2000). Les voyelles donnent du volume au message parlé car elles contiennent la plus grande partie de l'énergie sonore, les consonnes permettent, quant à elles, la compréhension et la distinction de l'information émise par la voix (Pulakka, 2012).

4.5. Effet du type de voix

Le contenu fréquentiel, notamment en HF, diffère selon le type de voix. La détection des HF semble déprendre du timbre de la voix, qui est spécifique à l'individu (Monson, 2011). Par conséquence, les HF varient qu'il s'agisse d'une voix modale, d'une voix soufflée ou d'une voix pressée.

Des études ont montré que les voix soufflées ont un niveau de HF plus élevé que les voix modales (Shoji et al., 1992; Valencia et al., 1994 ; Monson et al., 2012b). Par exemple, l'étude de Shoji et al. (1992) démontre au travers d'une comparaison de voix modales féminines et de voix soufflées féminines, produisant le phonème [a], qu'il existe une différence significative en termes de fréquences.

En résumé, les hautes fréquences ont longuement été ignorées. Les hautes fréquences jouent un rôle important dans l'aspect naturel de la parole ainsi que dans la localisation et la directivité de celle-ci. Ceci explique le nouvel intérêt dans la littérature scientifique pour les hautes fréquences. De plus, elles offrent des informations quant au genre du locuteur, mais également une meilleure intelligibilité de la parole et un meilleur développement du langage.

Chapitre 3 - La synthèse de parole

1. Définition

La synthèse de la parole est un système permettant de créer des stimuli de parole de façon artificielle. Cela se fait entre autres par le biais de l'informatique.

La synthèse de la parole est omniprésente dans notre société. Elle l'est dans des applications de systèmes de navigations, dans des applications de livres audio, dans les annonces de transports en commun, ou encore dans les assistants numériques (notamment pour la communication alternative et augmentée et pour les assistants numériques personnels tels qu'en matière de domotiques, de smartphones ou encore d'automobiles) (Wagner, 2019). Un exemple parlant est sans doute la voix synthétique qui a accompagné Stephen Hawking durant une grande partie de sa vie afin de lui permettre de communiquer (Wagner, 2019). À l'heure actuelle, ces voix de synthèses restent peu naturelles, pourtant, cet aspect de la naturalité est un élément essentiel pour les méthodes de synthèses et nécessite une nette amélioration (Gully, 2017).

En pratique, il existe deux grandes catégories de synthèse de la parole. Les méthodes non-paramétriques et les méthodes paramétriques que nous allons détailler ci-dessous.

2. Synthèses non-paramétriques

Les méthodes de synthèses non-paramétriques ne permettent pas de manipuler les caractéristiques prosodiques[9] de la source vocale et du conduit vocal indépendamment l'une de l'autre (Birkholz et al., 2017). Parmi les méthodes de synthèses non-paramétriques, nous détaillerons ci-dessous la synthèse concaténative et la synthèse par apprentissage profond.

9 Les caractéristiques prosodiques font référence à l'intonation, le débit de parole, l'expression, le rythme, l'accentuation d'un accent.

2.1. Synthèse concaténative

La synthèse concaténative est une méthode de synthèse vocale apparue dans les années 1990 (Birkholz, 2013). Grâce à l'avancée de l'informatique, il a été possible d'enregistrer des personnes et de reproduire les sons souhaités. Le principe général de cette méthode est de coller bout à bout des morceaux d'enregistrements tout en manipulant leur hauteur, leur durée et leur intensité afin qu'ils se correspondent. Cette méthode permet également d'analyser et de quantifier facilement les enregistrements de la parole naturelle (Birkholz et al., 2017). Cependant, la méthode est limitée à l'enregistrement qui a été réalisé, ce qui offre parfois une inintelligibilité de certains sons (Remi Blandin, communication personnelle, le 9 mars 2021).

La synthèse concaténative permet de manipuler facilement les caractéristiques prosodiques dites primaires, telles que la hauteur, la durée, et l'intensité (Birkholz et al., 2017 ; Hunt & Black, 1996). Néanmoins, la manipulation des caractéristiques prosodiques secondaires, qui interviennent dans la qualité de la voix, telles que la longueur du conduit vocal, la précision articulatoire, la nasalité et le type de phonation, demeure compliquée (Birkholz et al., 2017). Dans leur étude, Birkholz et al., (2017) expliquent que manipuler les caractéristiques prosodiques secondaires nécessite des actions très spécifiques au niveau articulatoire, comme par exemple une modification de la localisation du velum.

2.2. Synthèse par apprentissage profond

L'apprentissage profond ou « *Deep Learning* » est une forme d'intelligence artificielle. Il est basé sur un réseau de neurones artificiels qui s'inspire du cerveau humain. Sous ce réseau se cachent des centaines de couches de neurones qui permettent d'analyser les données petit à petit au fil du passage par chacune des couches. Au fil des transformations linéaires, des transformations non-linéaires et suite à l'association de chaque information, le système est capable de résoudre des taches complètes. Par exemple, pour lire une phrase, le système commencera par discriminer et reconnaitre chaque lettre, il s'attèlera ensuite aux mots avant de se concentrer sur la phrase. De nombreuses informations sont introduites dans le système qui les stocke et les associe ensuite comme le ferait notre cerveau par le biais de ses neurones. Dans ce système informatique, cela se fait par le biais d'algorithmes (Ben Aissa, 2020).

La synthèse par apprentissage profond est la méthode de synthèse vocale utilisée par le système *Siri d'Apple* ou encore la synthèse de la voix (l'assistant *Alexa d'Amazon*) (Briot, 2019). Cependant, elle demande énormément de puissance informatique et ne nous apprend aucune information sur la parole, ni sur comment cela se déroule pour produire la parole (Rémi Blandin, communication personnelle, le 9 mars 2021).

3. Synthèses paramétriques

Les méthodes de synthèses paramétriques permettent de manipuler les caractéristiques prosodiques de la source vocale et du conduit vocal indépendamment l'une de l'autre (Birkholz et al., 2017). Parmi les méthodes paramétriques, nous détaillerons ci-dessous la synthèse par règles et la synthèse articulatoire.

3.1. Synthèse par règles

La synthèse par règles, également appelée la synthèse par formants est une méthode qui permet la manipulation des caractéristiques prosodiques secondaires comme la longueur du conduit vocal, la précision articulatoire, la nasalité et le type de phonation au niveau du mot (Birkholz et al., 2017). Pour ce faire, des règles sont appliquées pour modéliser les formants par le biais d'un contrôle de leurs fréquences, de leurs amplitudes et également des caractéristiques des plis vocaux (Huang, 2001). Cette méthode modélise la parole dans le domaine spectral (Birkholz et al., 2017). Pour aboutir à ce type de synthèse, les chercheurs ont analysé des spectrogrammes afin de produire les mouvements qui étaient réalisés pour tel phonème à telle fréquence (Remi Blandin, communication personnelle, le 9 mars 2021). Cette méthode a été, pendant plusieurs décennies, le premier choix pour la synthèse et l'analyse des caractéristiques prosodiques secondaires (Birkholz et al., 2017).

3.2. Synthèse articulatoire

Parmi les différentes méthodes de synthèses paramétriques, la synthèse articulatoire est la seule à modéliser la parole sur base du domaine articulatoire (Birkholz et al., 2017). En effet, la synthèse articulatoire offre la possibilité de manipuler les caractéristiques prosodiques de la

source vocale et du conduit vocal (Birkholz et al., 2017). Cette méthode permet de modifier toutes les caractéristiques prosodiques directement au niveau articulatoire et physiologique (Birkholz et al., 2017). Le but de cette méthode réside dans la simulation de mouvements articulatoires proche des mouvements réels car cela est indispensable pour créer une parole se rapprochant le plus du naturel tout en étant intelligible (Birkholz, 2013).

Comme son nom l'indique, la synthèse articulatoire contrôle principalement le paramètre des articulateurs. Ce paramètre regroupe des éléments tels que l'ouverture labiale, la position et la hauteur de la langue, la position et la hauteur de l'apex lingual (Kröger, 1992). En modulant des paramètres comme les articulateurs, les chercheurs visent à définir le lien qui existerait entre les changements physiologiques et les sons produits. En d'autres termes, les chercheurs essayent d'appréhender l'impact des modifications géométriques du conduit vocal lors de la phonation (Huang, 2001). Cette méthode de synthèse vise à créer un signal de parole à travers une simulation numérique du passage du flux d'air dans le conduit vocal, en intégrant des dimensions neuromusculaires, articulatoires, aérodynamiques, acoustiques (Huang, 2001 ; Scully, 1990) mais également l'influence de la physionomie du conduit vocal sur les signaux laryngés (Huang, 2001).

La synthèse articulatoire est une technique complexe car elle nécessite de modéliser énormément d'interaction et de coarticulation. En effet, la synthèse articulatoire requiert une interaction entre le modèle de tractus vocal, les plis vocaux et le contrôle articulatoire (Birkholz, 2013). La qualité de la synthèse vocale dépend du réalisme de chaque modèle individuel (Birkholz, 2013). De plus, le contexte influence fortement la coarticulation de la synthèse articulatoire. Birkholz (2013) prend l'exemple de la consonne /g/ qui n'est pas prononcée de la même façon selon son contexte : *«* / gu / est articulée avec un corps de langue plus rétracté et avec des lèvres plus arrondies que dans la syllabe / gi / *»* (Birkholz, 2013, p. 1). Plusieurs études ont démontré que le contexte joue un rôle dans la coarticulation de la synthèse vocale, notamment dans la production et la perception de la parole synthétisée (Birkholz, 2013). En effet, la qualité de l'articulation et de l'acoustique de la coarticulation est issue de l'interaction dynamique des mouvements (Fowler & Saltzman, 1993). En 2013, Birkholz explique que dans l'énoncé /da/, nous réalisons un geste pour la réalisation de /d/ et un geste pour articuler la voyelle /a/. Ces derniers se font simultanément et se chevauchent. Ils sont en quelque sorte en compétition pour le contrôle du corps de la langue et de la mâchoire des articulateurs partagés. Lorsque nous prononçons un tel énoncé, une interaction de système dynamiques a lieu de sorte

que les gestes concurrents sur les articulateurs se mélangent. Cette découverte est récente, en effet, il y a quelques temps, les chercheurs se concentraient sur la simulation de phénomène articulatoire, sans se focaliser sur la perception et/ou sur la qualité de la synthèse vocale articulatoire (Birkholz, 2013).

Dans ce mémoire, nos stimuli seront des voix de synthèse articulatoire. Plus précisément, nos stimuli seront générés au travers de trois degrés de réalismes physiques : le modèle acoustique tridimensionnel (3D), le modèle acoustique unidimensionnel (1D) et l'algorithme d'extension de bande (BWE). Nous allons les détailler ci-dessous.

3.2.1. Modèle acoustique tridimensionnel (3D)

Un modèle acoustique tridimensionnel (3D) du conduit vocal a été développé pour représenter la forme variable dans le temps des voies respiratoires supra-glottiques (Birkholz, 2013). Ces innovations sont rendues possibles grâce à l'imagerie par résonnance magnétique (IRM), permettant de construire des modèles détaillés du conduit vocal et de l'articulation (Birkholz, 2013). Ce modèle de propagation acoustique simule correctement les HF, alors que le précédent modèle 1D n'intégrait lui que les basses fréquences (Arnela et al., 2019). Cette forme 3D du modèle articulatoire est la base d'un calcul précis des fonctions de surface du conduit vocal pour la simulation acoustique. Les modèles acoustiques 3D permettent la simulation numérique de voyelles, de diphtongues et de certaines séquences voyelle-consonne-voyelle en utilisant des géométries réalistes du conduit vocal (Freixes, 2021).

L'utilisation du modèle 3D a déjà été étudié dans une expérience antérieure, dans le cadre d'un mémoire qui précède le nôtre. En effet, Hoonaert (2021) a cherché à démontrer si le modèle 3D offrait des stimuli plus naturels que le modèle 1D dont nous parlerons dans le point suivant. Pour ce faire, elle a évalué 31 participants au travers de deux tâches perceptives. La première était une tâche de discrimination de paires de stimuli où les juges devaient indiquer si la paire de stimuli était identique ou différente. La seconde tâche était une tâche de perception où les participants entendaient les phonèmes un à un et devaient évaluer leur degré de naturel sur une échelle de Likert allant de 0 « pas du tout naturel » à 3 « totalement naturel ». Les résultats de cette étude montrent un effet significatif pour la perception des différences entre les paires de

stimuli entre le modèle 1D et le modèle 3D, mais aucun effet significatif n'a été démontré pour la seconde tâche. L'aspect naturel semblait dépendre du phonème.

3.2.2. Modèle acoustique unidimensionnel (1D)

Le modèle acoustique unidimensionnel (1D) est un modèle selon lequel le champ acoustique est le résultat de la propagation d'ondes planes (Sondhi & Schroeter, 1987). Nous pouvons définir une onde plane comme étant « une unique variation de pression acoustique uniforme sur une section transverse du tractus vocal » (Hoornaert, 2021, p. 26). Cela signifie que l'amplitude des ondes planes ne dépend pas de la courbature et de la forme de la partie transverse du tractus vocal, car elle dépend uniquement de l'aire de cette partie transverse (Hoonaert, 2021). C'est pourquoi, cette unique dépendance à l'aire de la section transverse représente une modèle unidimensionnel. Ce modèle est fréquemment utilisé en raison de sa simplicité. Toutefois, les ondes planes ne peuvent par décrire des fréquences supérieurs à 5 kHz et se cantonnent donc aux basses fréquences (Arnela et al., 2019 ; Freixes et al., 2018 ; Freixes et al., 2019).

3.2.3. Algorithme d'extension de bande (BWE)

À l'heure actuelle, de plus en plus d'appareils technologiques prennent en charge des applications ou des systèmes de communications vocales de haute qualité. Cela induit des bandes passantes très larges (Bachhav et al., 2018a). Malheureusement, tous les appareils ne prennent pas en charge ces bandes passantes et sont limités à des bandes passantes étroites. Une réduction de la bande passante s'accompagne d'une réduction de la qualité de la parole. Pour pallier cela, et obtenir, tout de même, *in fine* des communications de haute qualité, il est possible d'utiliser l'extension artificielle de la bande passante (Bachhav et al., 2018b).

Cette extension se fait par le biais d'algorithmes. Celui-ci se base sur un « modèle de source filtre classique dans lequel les informations d'enveloppe spectrale et d'erreur résiduelle sont extraites d'un signal à large bande en utilisant une analyse de prédiction linéaire conventionnelle » (Bachhav et al., 2018a, p. 1). Une sorte de « miroir spectrale est ensuite utilisée pour étendre la composante d'erreur résiduelle avant qu'un signal à très large bande étendu ne soit dérivé de sa combinaison avec l'enveloppe à large bande d'origine » (Bachhav et al., 2018a, p. 1).

L’extension artificielle de la bande passante cible principalement l'extension des signaux vocaux NB (narrow band = bande étroite) aux signaux vocaux WB (wide band = large bande). Le but de ces algorithmes est d’améliorer la qualité de la parole en cas d’utilisation d’appareil à large bande (de 50 Hz à 7 KHz) avec des appareils ou une infrastructure à bande étroite (de 0.3 kHz à 3.4 kHz) (Bachhav et al., 2018b). Dans ces cas, il existe un potentiel substantiel d'amélioration de la qualité : les composantes vocales significatives entre la limite NB de 4 kHz et la limite WB de 8 kHz peuvent être récupérées de manière fiable à l'aide de l'algorithme d'extension de bande passante artificielle (Bachhav et al., 2018a ; Bachhav et al., 2018b).

En réalité, l’extension de bande n’est pas une méthode spécifique ni à la synthèse, ni à la synthèse articulatoire. L’extension de bande est une méthode très approximative car elle prend en compte uniquement les paramètres du signal de la bande limité, et sur base de ces paramètres, elle construit ce qu’il pourrait exister sur une bande plus étendue (R. Blandin, communication personnelle, 27 avril 2022). C’est pourquoi, dans le cadre de ce mémoire, nous nous attendons à ce que les stimuli générés à partir de cette méthode soient considérés comme étant moins naturels que les stimuli générés par le modèle acoustique 1D et le modèle acoustique 3D. Nous y reviendrons dans la suite de ce travail (III, chap. 2).

3.2.4. Avantages et inconvénients de la synthèse articulatoire

Les avantages de la synthèse articulatoire sont nombreux. Tout d’abord, elle permet d’aboutir à un signal réaliste (Huang, 2001). Les signaux provenant d’un large panel de locuteur peuvent être synthétisés, peu importe leur genre (homme/femme), la fréquence de leur voix (aigue/grave), leur âge (enfant/adulte) (Shadle & Damper, 2001). En effet, la synthèse articulatoire a l’avantage de ne pas être influencée par des effets liés au locuteur (Birkholz et al., 2017 ; Delvaux & Pillot- Loiseau, 2019). En outre, elle permet de personnaliser le signal vocal par le contrôle des articulateurs (Huang, 2001).

Malgré les avantages évoqués ci-avant, il existe tout de même certains inconvénients liés à la synthèse articulatoire. Il n’est pas toujours aisé de trouver un équilibre entre la facilité de contrôle et la précision lors de la construction du modèle articulatoire (Tabet & Boughazi, 2011). De plus, cette méthode nécessite énormément de temps de travail (Birkholz et al., 2017).

Dans le point suivant (II, chap. 3, 4), nous allons présenter le logiciel de synthèse articulatoire qui a été utilisé pour réaliser les stimuli sonores de nos deux expériences.

4. Logiciel VocalTractLab

VocalTractLab[10] a été créé par Peter Birkholz et son équipe. Il s'agit d'un logiciel proposant un synthétiseur articulatoire de la parole. Il comprend un modèle géométrique 3D détaillé du conduit vocal qu'il est possible de contrôler par la manipulation de la forme et de la position de 23 paramètres vocaux tels que la protrusion labiale, la position linguale, l'ouverture glottique, la tension des plis vocaux, l'angle d'ouverture de la mâchoire, la position du voile du palais, l'élévation linguale, etc. (voir *Figure 7*) (Birkholz, 2013). La représentation de ces divers éléments contribue à des énoncés de haute qualité acoustique et à une meilleure compréhension des processus de production de la parole (Birkholz et al., 2006). Le logiciel crée chaque énoncé en partant de zéro. Pour ce faire, une chronologie de gestes articulatoires sont traduits en trajectoires d'articulateurs de la parole dans un conduit vocal en 3D. Ensuite, le logiciel génère des signaux vocaux sur base d'une simulation aérodynamique-acoustique (Lasarcyk, 2013).

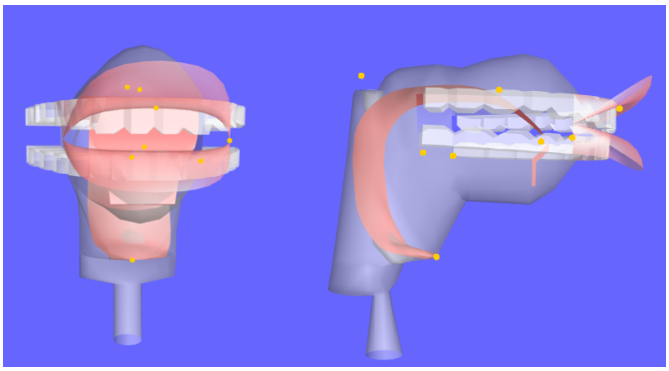

Figure 7 : Captures d'écran prises personnellement sur le site VocalTractLab le 3 juin 2021

Elles montrent les paramètres articulatoires modifiables sur une construction de face (à gauche) et de profil (à droite).

Des paramètres permettent aussi de spécifier la pression sous-glottique, la fréquence fondamentale et la forme de repos de la glotte. En outre, il est possible de régler, comme l'illustre la *Figure 8*, la partition gestuelle sur base de 8 niveaux modifiant les types de gestes

10 https://www.vocaltractlab.de (Birkholz, P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis, PLoS ONE, 8 (4), doi:10.1371/journal.pone.0060603.)

articulatoires. Les types de gestes pouvant être modifiés sont représentés par des points jaunes sur la *Figure 7* (ex : les gestes des lèvres, de la pointe de la langue, …) (Birkholz et al., 2017).

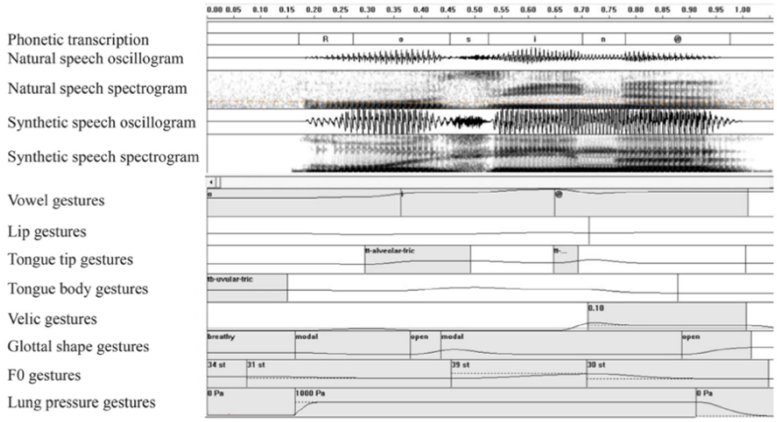

Figure 8 : Éditeur de parution gestuelle.

La partie supérieure montre les formes d'onde et les spectrogrammes des énoncés naturels et synthétiques du mot « Rosine ». La partie inférieure, elle permet la synthétisation de la parole articulatoire (Birkholz et al., 2017).

Un des avantages du logiciel est l'importance de la bande passante des stimuli (de 0.002 à 20 kHz). Elle permet à l'expérimentateur de couvrir l'ensemble des fréquences audibles par l'homme. Il est donc possible de remplir le but de notre recherche et de tester le rôle des HFE.

En résumé, bien qu'il existe différents types de synthèse, la synthèse articulatoire semble être celle qui offre un signal de parole de haute qualité grâce à la modélisation physique de la production de la parole. Les nombreux paramètres qu'il est possible de moduler permettent, quant à eux, une grande flexibilité. Elle permet de réaliser des mouvements articulatoires proche de l'appareil phonatoire humain, elle fournit donc une parole plus naturelle, de haute qualité, et intelligible. Dans la synthèse articulatoire, le modèle acoustique 3D tient compte de l'énergie acoustique des hautes fréquences. Il est le modèle qui parait le plus encourageant pour favoriser l'aspect naturel de la parole des stimuli synthétiques.

III. OBJECTIFS ET HYPOTHÈSES

Chapitre 1 - Objectifs

Les avancées technologiques permettent aujourd'hui de synthétiser et de reproduire la parole humaine. Toutefois, l'aspect naturel de cette parole de synthèse demeure un défi. Certains chercheurs postulent qu'une piste pour réduire ce différentiel d'aspect naturel entre la parole de synthèse et la parole humaine se situe au niveau des HF. Lors de ce travail, nous avons essayé de déterminer le rôle des HF dans l'aspect naturel de la parole de synthèse, à l'aide de deux tâches perceptives. Notre question générale est la suivante : « ***En synthèse articulatoire, comment les divers degrés de réalismes physiques dans la génération des hautes fréquences impactent-ils la perception de l'aspect naturel de la parole chez les jeunes adultes ?*** ». Pour répondre à cette question, nous allons recourir à trois méthodes permettant de générer les HF : (1) un modèle acoustique unidimensionnel (1D), (2) un modèle acoustique tridimensionnel (3D), et (3) un algorithme d'extension de bande (BWE).

Nous réaliserons deux expériences perceptives :

- La première a pour but de comparer l'aspect naturel de la parole obtenu avec les différentes méthodes. Dans un paradigme de comparaison par paires, les stimuli générés avec chacun des modèles seront comparés deux à deux.
- La seconde a pour but d'évaluer l'aspect naturel de chacun des stimuli générés avec chacun des trois modèles, à l'aide d'une échelle métrique.

Chapitre 2 - Hypothèses (H) et questions de recherche (QR)

Plus la méthode de synthèse est réaliste du point de vue de la physique acoustique, plus les stimuli générés devraient être perçus comme étant naturels. Cela implique que :

- (H1) : Les stimuli générés par le modèle 3D seront perçus comme plus naturels que les stimuli générés par le modèle 1D.
- (H2) : Les stimuli générés par le modèle 3D seront perçus comme plus naturels que les stimuli générés par le modèle BWE.
- (H3) : Les stimuli générés par le modèle 1D seront perçus comme plus naturels que les stimuli générés par le modèle BWE.

Ces trois hypothèses seront testées dans la première et dans la seconde tâche expérimentale. En plus de ces hypothèses générales, nous explorerons les questions de recherche suivantes :

- (QR 1) : « ***Est-ce que les divers degrés de réalisme dans la génération des hautes fréquences impactent la perception de l'aspect naturel différemment pour les stimuli masculins VS féminins ?*** ». Cela sera testé dans la seconde tâche expérimentale.

- (QR 2) : ***« Est-ce que les divers degrés de réalisme dans la génération des hautes fréquences impactent la perception de l'aspect naturel différemment pour les stimuli de voix pressée VS voix modale ?*** ». Cela sera testé dans la seconde tâche expérimentale.

- (QR 3) : « ***Est-ce que les divers degrés de réalisme dans la génération des hautes fréquences impactent la perception de l'aspect naturel différemment pour les phonèmes [a], [e], [i], [o], [u] ?*** ». Cela sera testé dans la première et dans la seconde tâche expérimentale.

- (QR 4) : « ***Les réponses des différents participants seront-elles équivalentes ?*** **».** Cela revient à tester la fiabilité inter-juges. Cela sera testé dans la seconde tâche expérimentale.

IV. MÉTHODOLOGIE

Chapitre 1- Sélection des participants

1. Comité d'éthique

Avant de réaliser cette étude, nous avons déposé une demande au comité d'éthique de la Faculté de Psychologie, Logopédie et des Sciences de l'Éducation (FPLSE). Le comité d'éthique nous a délivré un avis favorable le 13/11/2021 (numéro de dossier : 4820).

2. Recrutement

La procédure de recrutement des participants s'est réalisée sur base du bouche à oreille et via les réseaux sociaux, notamment sur base d'une annonce sur le réseau social « *Facebook* ». Les différents moyens de recrutement gardaient une neutralité de notre part afin de ne pas toucher à la liberté des participants d'accepter ou non de participer à notre étude.

3. Critères d'inclusion et d'exclusion des participants

Lorsque nous avons rencontré les participants, qui se sont désignés sur base de volontariat, nous leur avons demandé de remplir un **questionnaire anamnestique** (*Annexe 1*). Ce dernier a permis d'assurer l'homogénéité des participants, et de vérifier s'ils rentraient bien dans les critères d'inclusion (âge, langue maternelle, degré d'expertise, critères d'exclusions). Cependant, notre projet n'avait pas pour but d'établir un diagnostic, nous avions donc décidé de réaliser la totalité des expériences sur chaque participant. Et ce, même si nous nous rendions compte en cours d'expérience que certains participants ne pouvaient pas être inclus dans les résultats (par exemple, en cas de perte auditive non diagnostiquée). De plus, les participants devaient nous fournir un consentement écrit sur base des détails relatifs à notre expérimentation que nous leur avions préalablement soumis.

3.1. Critères d'inclusion

Sur base de la littérature scientifique, nous avions décidé que la tranche d'**âge** des participants se situerait entre 18 et 25 ans, dans l'espoir d'avoir une plus grande sensibilité aux HF (Hunter, 2020). Nous ne souhaitions pas que les participants soient des enfants pour éviter les possibles inattentions ou les manques de concentrations dans des tâches de perception (Rodríguez-Valiente et al., 2014). Les jeunes adultes sont donc un parfait compromis étant donné qu'ils se situent aux prémices du déclin de la sensibilité aux HF tout en ne possédant pas le manque d'attention des enfants. Nous avions également décidé que la **langue maternelle** des participants serait le français. En effet, des études montrent qu'il existe un écart entre la perception de la langue maternelle ou la perception d'une langue non-maternelle (Terbeek, 1977). Au niveau du **degré d'expertise**, nous avons recruté des jeunes adultes actuellement étudiants à l'Université de Liège, peu importe la faculté. De la sorte, les participants sont des étudiants universitaires avec un *background* similaire, le fait qu'ils étudient à l'Université de Liège permettait la facilité d'accès au lieu du testing, qui se déroulait sur le campus du Sart-Tilman

3.2. Critères d'exclusion

Les participants ne devaient pas présenter d'antécédents auditifs ou de troubles auditifs. C'est pourquoi, une audiométrie tonale a été réalisée sur les fréquences de 0.5 kHz, 1 kHz, 2 kHz, 8 kHz, 12.5 kHz afin de nous assurer que les participants disposaient d'une bonne sensibilité auditive pour percevoir les HF durant l'expérience. Si cela n'était pas le cas, autrement dit, si les participants avaient une perte auditive, pour l'une des deux oreilles, supérieure à 20 dB dans des fréquences comprises entre 0.5 kHz et 12.5 kHz, alors, leurs résultats étaient exclus de notre étude. Cependant, cela n'a pas été le cas dans ce mémoire. Les résultats de l'audiométrie tonale de chaque participant sont présentés dans l'*Annexe 4*.

3.3. Description de l'échantillon final

	Total (n)	% calculé sur les 40 participants
Genre		
Masculin	20	50
Féminin	20	50
Âge		
19 ans	4	10
20 ans	2	5
21 ans	10	25
22 ans	8	20
23 ans	11	27,5
24 ans	5	12,5
Année d'étude		
B1	2	5
B2	3	7,5
B3	9	22,5
M0	1	2,5
M1	16	40
M2	9	22,5
Filière d'étude		
Logopédie	19	47,5
Psychologie	2	5
Sciences humaines et sociales	4	10
Sciences de l'éducation	1	2,5
Droit	5	12,5
Kinésithérapie	1	2,5
HEC	1	2,5
Ingénieur de gestion	2	5
Sociologie	2	5
Ingénieur civil	1	2,5
Sciences politiques	1	2,5
Médecine	1	2,5
Langue maternelle		
Français	40	100
Autre	0	0
Bilinguisme		
Oui	10	25
Non	30	75
Antécédents d'audition		
Oui	0	0
Non	40	100
Connaissance sur la synthèse de parole		
Oui	4	10
Non	36	90

Tableau 2 : Description de l'échantillon final

L'âge moyen de notre échantillon est de 21.875 ans quant à l'écart type de l'échantillon, il s'élève à 1.451. L'âge minimum est de 19 ans, quant à l'âge maximum, il est de 24 ans. Si nous distinguons selon le genre. Nous avons un âge moyen pour notre échantillon féminin de 22.05 ans, un écart type de 1.117, un âge minimum de 20 ans et maximum de 24 ans. Quant aux participants de la gente masculine, l'âge moyen de l'échantillon était de 21.6 ans, l'écart type de 1.625, l'âge minimum de 19 ans et maximum de 24 ans.

Seules 4 personnes au sein de notre échantillon possèdent des connaissances en matière de synthèse de la parole. Pour 3 d'entre elles, il s'agit de connaissances sommaires acquises dans le cadre de cours dispensés à l'université. Pour 1 personne, il s'agit de connaissances acquises individuellement lors de recherches menées sur le sujet dans le cadre d'un travail personnel réalisé sur l'influence de la musique sur le corps humain. Cet individu possède par conséquent des connaissances théoriques et pratiques en la matière.

Notre échantillon est composé de 10 personnes en situation de multilinguismes. 5 d'entre elles sont bilingues anglais-français, 3 d'entre elles sont trilingues anglais-français-néerlandais, 1 d'entre elles est trilingue anglais-arabe-français, et 1 d'entre elles est trilingue anglais-français-polonais.

Chapitre 2 - Description du matériel

1. Environnement de l'étude

Les testings ont été réalisés dans la cabine audiométrique du CEDIA, située dans le quartier Polytech du campus du Sart-Tilman à l'Université de Liège. Le CEDIA est un bureau d'études et de recherches spécialisé en acoustique et en vibration (http://www.cedia.ulg.ac.be). La cabine était meublée d'un bureau, de deux chaises, d'un haut-parleur XM6.D sn 07-3301 de la marque FAR by ATD, d'une souris portable et d'un tapis de souris. Nous y avons ajouté l'ordinateur ACER (intel core i3 8th gen) de l'Unité Logopédique de la Voix (ULV) pour réaliser les deux tâches perceptives. Le participant était invité à s'asseoir sur la chaise face à l'ordinateur pour réaliser les deux tâches perceptives (*Figure 9*).

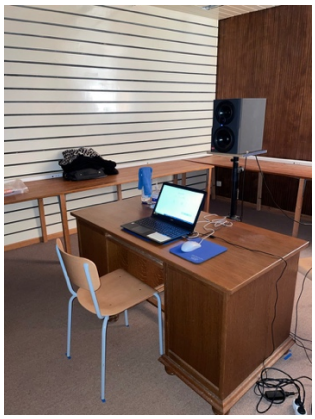

Figure 9 : Environnement de l'étude

Avant de débuter les testings, Monsieur Xavier Kaiser, ingénieur de recherches à l'Université de Liège, a réalisé une mesure de bruit de fond du local à l'aide du sonomètre 01dB type Fusion sn 10601, calibré avec un calibreur 01dB CAL21 sn 35293309. Le bruit de fond était de 25 dB SPL dans la cabine audiométrique.

Pour réaliser les deux tâches perceptives, le participant était assis face à un ordinateur portable. Les différents stimuli synthétiques étaient délivrés à 70 dB SPL via un haut-parleur qui se trouvait face au participant, à 1 mètre de distance (Monson, 2011 ; Monson, 2014 ; Hunter et al., 2012b ; Monson & Caravello, 2019). En effet, les HF se propagent de façon directionnelle, de la source vers le locuteur (Hunter, 2020). C'est pourquoi, nous avons fait preuve d'une grande attention en ce qui concerne l'emplacement du haut-parleur, afin de contrôler les effets de directivité et de diffusion des HF (Monson, 2014). De plus, le haut-parleur a été choisi en fonction de sa bonne amplification des HF (Monson, 2011).

L'éclairage était assuré par des néons, mais ceux-ci étant beaucoup trop bruyants, seule une lampe sur pied était allumée dans le coin de la cabine lors des testing, permettant ainsi un éclairage suffisant et un bruit de fond moins élevé (*Figure 10*). Toujours pour réduire le bruit de fond, nous avons pris soin de couper le chauffage lors de chaque testing, bien que parfois un bruit continu de souffle était émis.

Figure 10 : Éclairage de la cabine audiométrique

2. Audiométrie tonale

Nous avons réalisé une audiométrie tonale sur les 40 participants pour évaluer les seuils auditifs de chacun. Les fréquences 0.5 kHz, 1 kHz, 2 kHz, 4 kHz, 8 kHz et 12.5 kHz ont été testées avec l'audiomètre MADSEN Itera II (Voir *Figure 11*) et le casque Sennheiser HDA 300 (voir *Figure 12*). L'audiomètre a été calibré le 10/09/2021 (*Annexe 2*). Comme expliqué *supra* (IV, chap. 1, 3.2.), nous avions décidé d'exclure les participants avec une perte d'audition supérieure à 20 dB, afin d'éviter qu'ils présentent une déficience auditive, même légère. Dans ce mémoire, aucun des participants n'a dû être exclu de l'étude.

Figure 11 : Audiomètre MADSEN Itera II

Figure 12 : Casque Sennheiser HDA 300

3. Stimuli sonores utilisés

Notre expérience s'est basée sur des stimuli auditifs produits à l'aide de la synthèse articulatoire, cette méthode permettant de générer des phonèmes via le contrôle des paramètres des articulateurs (ouverture labiale, position de la hauteur de la langue, …). Les phonèmes qui ont été utilisés dans notre expérience ont été créés avec le logiciel VocalTractLab.

Phonèmes	Genre du locuteur	Qualité vocale
/a/	Homme	Modal
	Homme	Pressé
	Femme	Modal
	Femme	Pressée
/e/	Homme	Modal
	Homme	Pressé
	Femme	Modal
	Femme	Pressé
/i/	Homme	Modal
	Homme	Pressé
	Femme	Modal
	Femme	Pressé
/o/	Homme	Modal
	Homme	Pressé
	Femme	Modal
	Femme	Pressé
/u/	Homme	Modal
	Homme	Pressé
	Femme	Modal
	Femme	Pressé

Tableau 3 : Les phonèmes des deux taches expérimentales

Pour les deux tâches expérimentales, nous avons utilisé 5 phonèmes synthétiques, à savoir les voyelles : [a], [e], [i], [o], [u]. Le tableau 3 ci-dessus est un récapitulatif présentant la totalité des stimuli synthétisés avec chacune des 3 méthodes. VocalTractLab dispose de paramètres prédéfinis pour des phonèmes allemands (Rémi Blandin, communication personnelle, le 25 mai

2022), c'est pourquoi, nous avons souhaité nous concentrer sur ces 5 phonèmes uniquement car ce sont des voyelles à la fois présentes en français, en allemand et dans de nombreuses autres langues (Angélique Remacle & Rémi Blandin, communication personnelle, le 4 mai 2021). De plus, les voyelles correspondent aux sons de la parole qui possèdent des caractéristiques stables dans le temps lorsque celles-ci sont prononcées individuellement (Moore, 2003). Les stimuli que nous soumettrons aux participants seront limités à 10 kHz. Comme démontré précédemment, la bande de parole à 10 kHz semble plus naturelle (Moore, 2003).

Ces 5 phonèmes produits à la fois par une voix de synthèse féminine, et une voix masculine, ont été réalisé à travers 3 méthodes de synthèse différentes : 1) un modèle unidimensionnel du conduit vocal (1D), 2) un modèle tridimensionnel du conduit vocal (3D), et 3) une méthode d'extension de bandes (BWE). Pour chaque méthode, un total de 20 stimuli seront générés : 5 phonèmes ([a], [e], [i], [o], [u]) x 2 genres (homme, femme) x 2 types de voix (modale, pressée). Ces stimuli ont été synthétisés par Rémi Blandin et son équipe grâce au synthétiseur de la parole articulatoire, VocalTractLab (Birkholz, 2013 ; www.vocaltractlab.de).

Chapitre 3- Procédure expérimentale

1. Déroulement général de l'étude

Le testing s'est déroulé en une seule rencontre avec chaque participant. En moyenne, 1 heure approximativement par participant a suffi pour réaliser l'ensemble des taches, allant de leur arrivée à leur départ (voir *Annexe 5*).

Concrètement, chaque testing s'est déroulé de la façon suivante : nous avions rendez-vous dans le hall de l'Institut Montéfiore avec le participant, nous l'invitions à respecter les gestes barrières et à désinfecter ses mains à l'aide du gel hydroalcoolique présent sur place. Ensuite, nous l'accompagnions jusqu'à la cabine audiométrique. Dans la cabine audiométrique, nous proposions au participant de retirer sa veste, mais lui demandions de garder son masque chirurgical. Ensuite, nous lui fournissions les documents administratifs et lui proposions de compléter et signer chacun de ces documents, tout en restant disponible pour répondre aux éventuelles questions. Nous lui demandions de mettre son téléphone et son ordinateur portable en mode avion, pour éviter les interférences. Nous commencions par réaliser l'audiométrie

tonale, et ensuite proposions de passer à la première tâche expérimentale. Étant donné que nous nous trouvions dans un espace clos, dépourvu de fenêtre et insonorisé, nous demandions au participant s'il désirait réaliser une pause de 5 minutes, durant laquelle il pouvait prendre l'air. Par la suite, la seconde tâche expérimentale était réalisée. La *Figure 13* ci-dessous illustre le déroulement de ce testing.

Figure 13 : Déroulement du testing

1.1. Première tâche perceptive

La première tâche expérimentale était une comparaison par paires. Le participant avait comme tâche d'écouter deux phonèmes et puis de répondre à la question suivante « *Lequel des deux sons vous semble le plus naturel ?* ». Pour répondre à cette question, le participant disposait d'un ordinateur et d'une souris à fil. Sur l'ordinateur il découvrait une interface (voir *Figure 14*) sur laquelle il était inscrit « *Son 1* » et « *Son 2* ». Le participant était amené à cliquer sur ces deux boutons pour entendre les différents stimuli. Il avait la possibilité d'écouter ces deux stimuli autant de fois qu'il le désirait, comme dans d'autres études (Vitela, 2015 ; Birkholz et al., 2017 ; Baumann, 2008). Lorsque son choix était réalisé, le participant devait cliquer sur « *Le son 1 est le plus naturel* » ou « *Le son 2 est le plus naturel* » en fonction de son avis. Il était ensuite invité à appuyer sur « *Paire suivante* ». Bien que le participant pouvait écouter les différents phonèmes autant de fois qu'il le souhaitait, il n'avait pas le droit de revenir en arrière une fois qu'il avait décidé de passer à la paire suivante. Toutes les paires de phonèmes ont été testées durant une seule et même session.

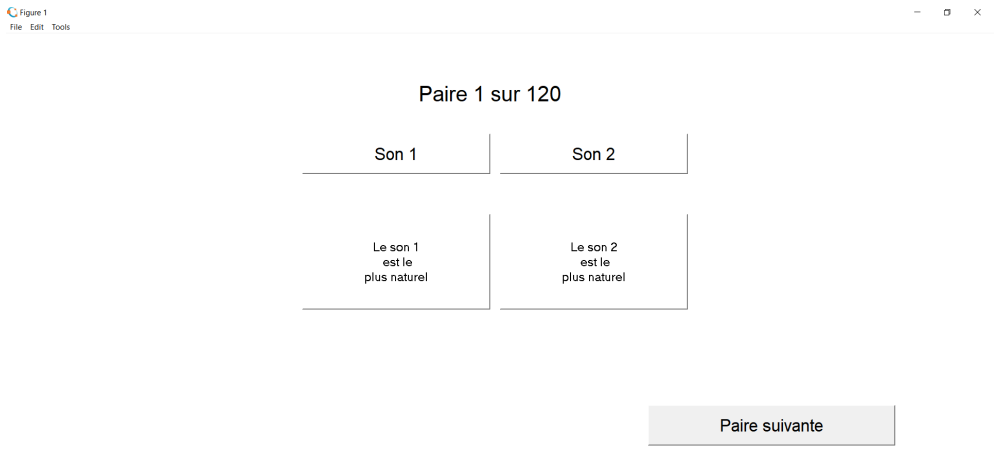

Figure 14 : Interface de la tache 1

La consigne suivante était donnée oralement : « *Vous allez écouter 2 sons, et vous devez indiquer quel est le son qui vous parait le plus naturel. Pour écouter les sons, vous devez cliquer sur le bouton « son 1 » ou sur le bouton « son 2 ». Ensuite vous choisirez le son qui vous parait le plus naturel en cliquant sur le bouton « le son 1 est plus naturel » ou « le son 2 est plus naturel ». Pour passer à la paire suivante, vous devez cliquer sur le bouton « paire suivante ». Vous pouvez écouter les sons autant de fois que vous le souhaitez. Néanmoins, lorsque vous aurez cliqué sur le bouton « paire suivante », vous ne pourrez plus revenir en arrière. Nous allons d'abord faire un entrainement avec 2 paires de sons. Durant le testing, vous aurez 120 paires de sons à comparer. Après l'essai, n'hésitez pas à poser vos éventuelles questions ».*

Afin que le participant se familiarise avec l'interface, il était amené à réaliser deux essais. Il s'agissait d'une phase d'entrainement où il entendait le phonème / ø/ en voix modale avec une voix de synthèse masculine pour les deux paires, mais avec des modèles différents (1D-3D-BWE). Chaque participant entendait les mêmes paires afin qu'ils possèdent tous les mêmes références. Le phonème évalué durant la phase d'entrainement était volontairement différent des 5 phonèmes utilisés dans le testing afin de ne pas déstabiliser le participant. Au total, la phase d'entrainement et la phase de test duraient en moyenne 25 minutes.

Nous avons décidé de réaliser une tâche de comparaison par paires sur base de la littérature scientifique. En effet, la fiabilité des résultats est meilleure si elle est réalisée sur base d'une comparaison par paires (Teston, 2004). En outre, cette comparaison par paires améliore la fiabilité inter-juges étant donné que comparer des paires de stimuli ne demandent pas aux participants de se rapporter à « leurs standards internes » (Kacha et al., 2005). C'est pourquoi,

ce choix nous semble le plus judicieux, dans le but d'avoir une bonne fiabilité des résultats. En ce qui concerne la fiabilité inter-juges, nous l'avons évaluée en administrant à chaque participant la même tache avec des stimuli identiques (Remacle et al., 2014). L'ensemble des données recueillies a automatiquement été enregistré dans un fichier csv.

1.2. Seconde tâche perceptive

La seconde tâche consistait en un jugement de l'aspect naturel des phonèmes. Plus précisément, le participant était amené à écouter les 120 mêmes stimuli présentés dans le tableau 3 *supra* (IV, chap. 2, 3), et répondre à la question « *Quel est le degré de naturel de ce son ?* ». Les phonèmes étaient présentés un à un. Comme le montre la *Figure 15*, l'interface était composée d'un bouton « *Écouter* » sur lequel le participant devait cliquer pour entendre le phonème. Comme pour la première expérience, il n'y avait aucune restriction quant au nombre d'expositions aux stimuli, ce qui signifie que le participant pouvait écouter chaque stimulus autant de fois qu'il le désirait, comme dans d'autres études (Vitela, 2015 ; Birkholz et al., 2017 ; Baumann, 2010). Le participant indiquait le degré de naturel à l'aide d'un curseur sur une échelle graduelle allant de 0 (pas du tout naturel) à 100 (totalement naturel). Toutefois, cette échelle métrique n'était pas totalement graduée, un marquage par dizaine était présent mais seuls les échelons extrêmes et centraux, autrement dit 0, 50 et 100 étaient indiqués (voir *Figure 15*). Le participant pouvait écouter le son autant de fois qu'il le voulait et déplacer le curseur autant de fois qu'il le souhaitait. Pour passer au son suivant, le participant devait cliquer sur « *Suivant* ». Une fois passé au son suivant, il ne pouvait plus revenir en arrière.

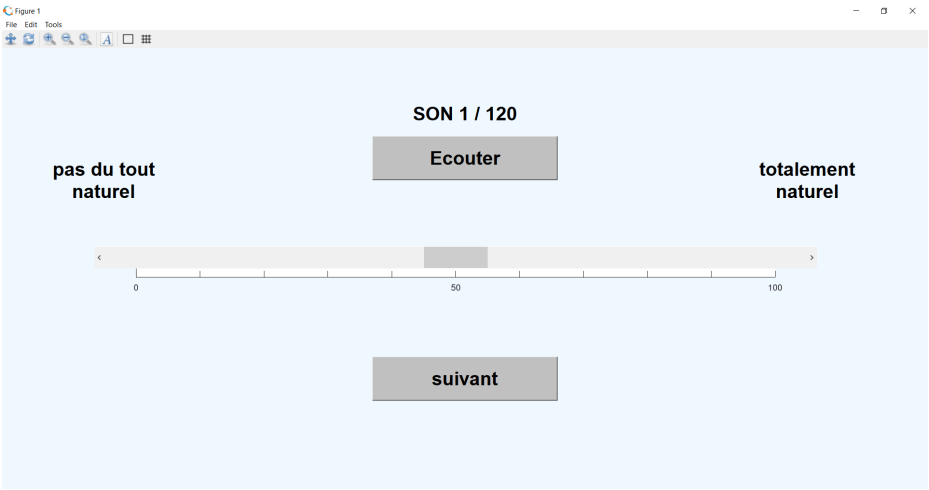

Figure 15 : Interface de la tache 2

La consigne était donnée oralement comme suit : « *Vous allez désormais écouter 120 sons, l'un à la suite de l'autre, et vous allez devoir évaluer leur degré de naturel sur base d'une échelle allant de 0 « pas du tout naturel » à 100 « tout à fait naturel ». Pas du tout naturel signifie que le son que vous écoutez ressemble à une voix artificielle, éloignée d'une voix réelle. Tout à fait naturel signifie que le son que vous écoutez ressemble à une voix réelle. L'écoute du son se réalise de la même façon que dans la tache 1. Quant au degré de naturel, vous le déterminerez en déplaçant le curseur se situant en bas de l'interface. Lorsque vous avez cliqué sur le bouton « paire suivante », le son suivant se lance automatiquement, mais vous pouvez tout de même écouter ce son autant de fois que vous le souhaitez en cliquant sur le bouton « écouter ». Néanmoins, une fois que vous avez cliqué sur le bouton « suivant », vous ne pouvez pas revenir en arrière. Nous allons réaliser un essai, après lequel vous pourrez poser vos éventuelles questions.* ».

Comme pour la première tâche expérimentale, le participant était amené à réaliser une phase d'entrainement pour se familiariser avec la tâche et avec l'interface. Le participant entendait le phonème / ø/ en voix modale avec une voix de synthèse masculine et le modèle 3D. Le 2ème essai était le phonème / ø/ en voix pressée avec une voix de synthèse masculine et le modèle d'algorithme d'extension. Le phonème évalué durant la phase d'entrainement était de nouveau volontairement différent des 5 phonèmes utilisés dans le testing afin de ne pas déstabiliser le participant. Au total, la phase d'entrainement et la phase de test duraient en moyenne 20 minutes. L'ensemble des données recueillies a automatiquement été enregistré dans un fichier csv.

Chapitre 4 - Analyses statistiques

L'ensemble des analyses statistiques de ce mémoire ont été accomplies par Vincent Didone, logisticien de recherche à l'Université de Liège et par nous-même. Grâce aux analyses statistiques, nous avons pu tester les différentes hypothèses et questions de recherches formulées (III, chap. 2) concernant la première et la seconde tâche expérimentale.

1. Analyses statistiques pour la première tâche expérimentale

Les analyses statistiques de la première expérience ont été **réalisées à l'aide du modèle de Bradley-Terry-Luce**. Classiquement, dans les expériences de comparaison par paires, ce modèle est utilisé car il n'existe pas d'autres possibilités d'analyser ce type de tâche. Ce modèle a été réalisé grâce au package *prefmod* et *gnm* dans le programme R pour des raisons inhérentes au plan expérimental. Ce package permet de réaliser des comparaisons, en utilisant la dernière condition par défaut comme étant la condition de référence. La présence d'un *odds ratio* permet de choisir la condition à comparer par rapport à une autre condition. Pour réaliser les différentes comparaisons, un réajustement du modèle a été réalisé afin de modifier la condition de référence.

Les variables indépendantes introduites dans le modèle sont les suivantes : le degré de réalisme physique du modèle acoustique (3D, 1D, BWE) et le type de phonème phonème ([a], [e], [i], [o], [u]). La variable dépendante est le jugement de naturalité des stimuli. Dans un premier temps, **un modèle par phonème** a été réalisé pour étudier la comparaison des conditions entre elles (soit 5 modèles au total, sans prise en compte de l'interaction entre les phonèmes). Dans un second temps, un **modèle général** a été réalisé sans considérer les autres variables indépendantes, telles que la qualité vocale, le genre de la voix de synthèse ou le type de phonème.

2. Analyses statistiques pour la seconde tâche expérimentale

Dans un premier temps, les analyses statistiques de la seconde expérience ont été réalisées à l'aide d'un **modèle linéaire mixte**, afin de contrôler l'effet aléatoire lié au participant. Ce modèle se traduit par une mise en interaction des différentes variables les unes avec les autres. Plusieurs variables indépendantes ont été intégrées, dont on a testé les **effets principaux** : le

degré de réalisme physique du modèle acoustique (3D, 1D, BWE), le genre de la voix de synthèse (féminin ou masculin), le type de phonème ([a], [e], [i], [o], [u]), la qualité vocale (modale ou pressée). L'effet de ces variables a été testé sur la variable dépendante : le score attribué au degré de naturalité du stimulus. De plus, plusieurs **interactions** entre les variables indépendantes ont été testées : degré de réalisme physique du modèle acoustique * type de phonème ; genre * type de phonème. Avant d'appliquer le modèle, la condition normalité a été vérifiée, et il s'avère qu'elle a été tolérée.

Dès qu'un effet significatif était présent, des **contrastes linéaires** ont été réalisés. Ils reposent sur la comparaison des différentes moyennes entre elles, afin de préciser où se situe précisément la différence lorsqu'il y a un effet. L'effet significatif des résultats a été interprété au moyen d'un test χ^2.

V. RÉSULTATS

1. Résultats des hypothèses (H) et questions de recherche de la première tâche expérimentale (QR)

1.1. Effet du degré de réalisme physique du modèle acoustique (H1, H2, H3)

Nous avons examiné l'influence du degré de réalisme physique sur le score attribué à l'aspect naturel des différents phonèmes. Pour rappel, nous avions émis trois hypothèses : **(H1) : Les stimuli générés par le modèle 3D seront perçus comme plus naturels que les stimuli générés par le modèle 1D ; (H2) : Les stimuli générés par le modèle 3D seront perçus comme plus naturels que les stimuli générés par le modèle BWE ; (H3) : Les stimuli générés par le modèle 1D seront perçus comme plus naturels que les stimuli générés par le modèle BWE**. Le modèle de Bradley-Terry-Lucce a permis deux comparaisons. Nous avons d'abord comparé le modèle 3D et le modèle BWE au modèle 1D. Ensuite, nous avons comparé le modèle BWE et le modèle 1D au modèle 3D. Les résultats figurent dans le tableau 4. Les résultats significatifs ($p < .05$) sont présentées en gras et est suivi d'astérisques. Un astérisque signifie que $p < .05$, deux astérisques signifient que $p < .01$ et trois astérisques signifient que $p < .001$.

	1D		**3D**		**BWE**	
	Score Z	**Probabilité (>|Z|)**	**Score Z**	**Probabilité (>|Z|)**	**Score Z**	**Probabilité (>|Z|)**
1D			5.017	**0.0001*** **	- 11.153	**0.0001*** **
3D					- 15.891	**0.0001*** **

Tableau 4 : Résultats des comparaisons réalisées entre les différents modèles acoustiques

Le tableau 4 montre que toutes les comparaisons des modèles deux à deux sont significatives. **Nous sommes en mesure de confirmer** les hypothèses **H1**, **H2**, et **H3**.

La *Figure 16* représente la valeur estimée de la force de la valeur[11] de chaque modèle sur une échelle allant de 0 à 1 par rapport à son jugement en termes de naturalité. Ce graphique a pour but de comparer des modèles acoustiques entre eux par rapport à des valeurs de références.

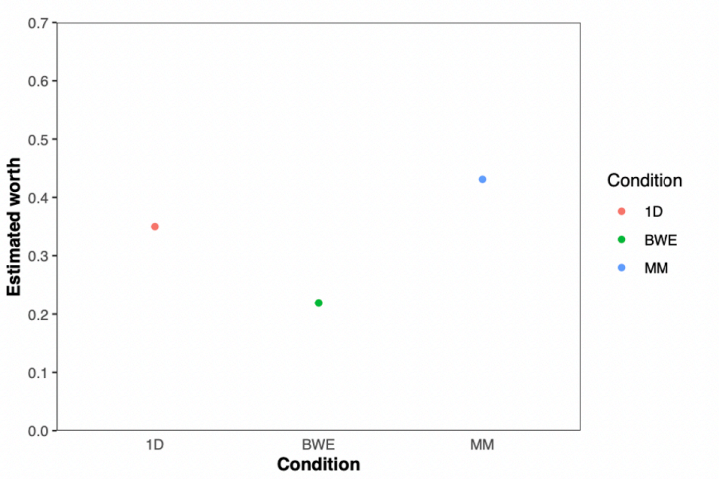

Figure 16 : Graphique de la force de valeur estimée des modèles acoustiques

Nous pouvons observer sur la *Figure 16* que le modèle 3D (MM) est statistiquement supérieur au modèle 1D, qui est lui-même supérieur au modèle BWE.

1.2. Effet du type de phonème (QR 3)

Nous avons évalué l'influence du type de phonème sur le score attribué à l'aspect naturel des phonèmes. Pour rappel, notre question de recherche était : « ***Est-ce que les divers degrés de réalisme dans la génération des hautes fréquences impactent la perception de l'aspect naturel différemment pour les phonèmes [a], [e], [i], [o], [u] ?*** ». Les analyses statistiques ont été réalisées avec le modèle de Bradley-Terry-Lucce, où deux comparaisons ont été effectuées. Dans un premier temps, le modèle 3D et le modèle BWE ont été comparés au modèle 1D. Dans un second temps, le modèle BWE et le modèle 1D ont été comparés au modèle 3D. Ce modèle ne permettant pas de comparer les phonèmes entre eux, nous allons présenter ci-dessous les analyses pour chaque phonème, indépendamment les uns des autres.

[11] La valeur estimée de la force de la valeur, aussi appelée « *Estimated worth parameters* » s'étale de 0 à 1. Il s'agit d'une sorte d'échelle qui permet de comparer les stimuli entre eux et de les classer correctement, indépendamment des mesures et/ou des signes négatifs. Lorsque l'on additionne les différents stimuli, les valeurs sont égales à 1.

En ce qui concerne le phonème [a], il existe une différence significative entre le modèle 1D et le modèle 3D ($Z = -2.868$, $p < .05$), ainsi qu'entre le modèle 1D et le modèle BWE ($Z = -12.415$, $p < .001$). Il existe également une différence significative entre le modèle 3D et le modèle BWE ($Z = -10.132$, $p < .001$). Concernant le phonème [e], il existe une différence significative entre le modèle 1D et le modèle BWE ($Z = -7.958$, $p < .001$), ainsi qu'entre le modèle 3D et le modèle BWE ($Z = -6.937$, $p < .001$). Il n'y a pas de différence significative entre le modèle 1D et 3D ($Z = 1.116$, $p > .05$). Pour le phonème [i], il existe une différence significative entre le modèle 1D et le modèle BWE ($Z = -9.094$, $p < .001$), ainsi qu'entre le modèle 3D et le modèle BWE ($Z = -8.682$, $p < .001$). Il n'existe pas de différence significative entre le modèle 1D et le modèle 3D ($Z = -0.470$, $p > .05$). En ce qui concerne le phonème [o], il existe une différence significative entre le modèle 1D et le modèle 3D ($Z = 6.433$, $p < .001$). Une différence significative est également présente entre le modèle 3D et le modèle BWE ($Z = -6.606$, $p < .001$). Il n'existe pas de différence significative entre le modèle 1D et le modèle BWE ($Z = -0.184$, $p > .05$). Pour le phonème [u], il existe une différence significative le modèle 1D et le modèle 3D ($Z = 8.712$, $p < .001$), également entre le modèle 1D et le modèle BWE ($Z = 5.636$, $p < .001$). Une différence significative est également objectivée entre le modèle 3D et le modèle BWE ($Z = -3.336$, $p < .01$). À la suite de ces différentes analyses, nous observons que **les divers degrés de réalismes dans la génération des hautes fréquences impactent la perception de l'aspect naturel différemment pour les phonèmes [a], [e], [i], [o], [u]**.

La *Figure 17* présente la valeur estimée de la force de la valeur de chaque stimulus sur une échelle allant de 0 à 1 par rapport à son jugement en termes de naturalité. Ce graphique permet une comparaison de stimuli entre eux par rapport à des valeurs de références. Cela permet de classer correctement les différents stimuli, indépendamment des mesures et indépendamment des signes négatifs ou non obtenus dans les analyses statistiques.

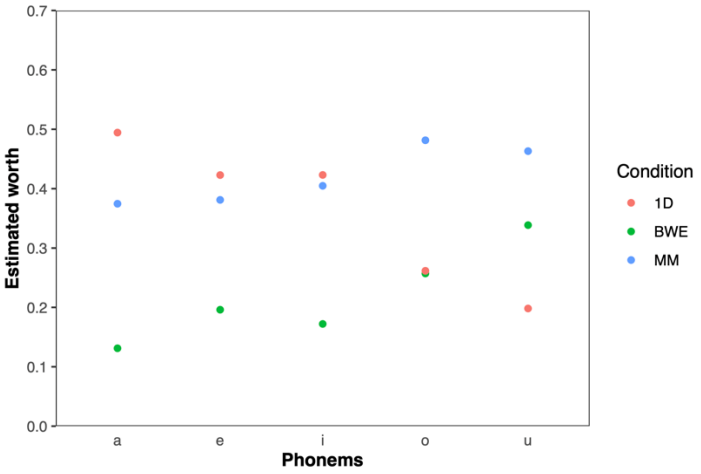

Figure 17 : Graphique de la force de la valeur estimée des phonèmes

La *Figure 17* illustre que les degrés de réalismes physiques des modèles acoustiques diffèrent en fonction du type de phonème. Pour les phonèmes [a] et [e], le modèle 1D est considéré comme étant le plus réaliste, suivi du modèle 3D, lui-même suivi du modèle BWE. Pour le phonème [i], les modèles 1D et 3D se placent au-dessus du classement, et ne sont pas jugés comme étant différents. Ils sont suivis du modèle BWE. Pour le phonème [o], le modèle 3D est jugé comme étant le plus réaliste, suivi du modèle 1D et BWE qui ne présentent pas de différence entre eux. Pour le phonème [u], le modèle 3D est considéré comme étant le plus réaliste, suivi du modèle BWE qui est lui-même suivi du modèle 1D.

2. Résultats des hypothèses (H) et questions de recherche (QR) de la seconde tâche expérimentale

2.1. Résultats des hypothèses et questions de recherches des effets principaux

Les résultats statistiques des hypothèses et questions de recherches qui concernent les effets principaux sont représentés dans le tableau 5. Lorsque des effets significatifs sont démontrés, une analyse plus approfondie de ces effets est réalisée au travers de contrastes linéaires, présentés dans les sections 2.1.1. à 2.1.4.

Variable indépendantes	Khi Carré (X^2)	Degrés de liberté (Df)	Probabilité (> X^2)
Degré du réalisme du modèle acoustique (H1, H2, H3)	16.604	2	**0.000*** **
Genre de la voix de synthèse (QR 1)	17.439	1	**0.000*** **
Type de phonème (QR 3)	14.792	4	**0.005** **
Qualité vocale (QR 2)	0.725	1	0.394

Tableau 5 : Résultats du modèle linéaire mixte

2.1.1. Effet du degré de réalisme du modèle acoustique (H1, H2, H3)

Pour rappel, nos hypothèses stipulaient que **(H1) les stimuli générés par le modèle 3D seront perçus comme plus naturels que les stimuli générés par le modèle 1D** ; **(H2) les stimuli générés par le modèle 3D seront perçus comme plus naturels que les stimuli générés par le modèle BWE** ; **et (H3) les stimuli générés par le modèle 1D seront perçus comme plus naturels que les stimuli générés par le modèle BWE**. Un effet significatif principal du modèle a été mis en avant (X^2 = 16.604; p < .001). Les contrastes linéaires indiquent que toutes les comparaisons deux à deux sont significatives. Comme illustré dans la *Figure 18*, le modèle 1D est considéré comme plus naturel que le modèle BWE (Z = 7.277, p < .001) ; le modèle 1D est considéré comme moins naturel que le modèle 3D (Z = - 5.579, p < .001) et le modèle BWE est considéré comme moins naturel que le modèle 3D (Z = - 12.829, p < .001). **Nous sommes en mesure de confirmer** nos hypothèses **(H1)**, **(H2)**, **(H3)**.

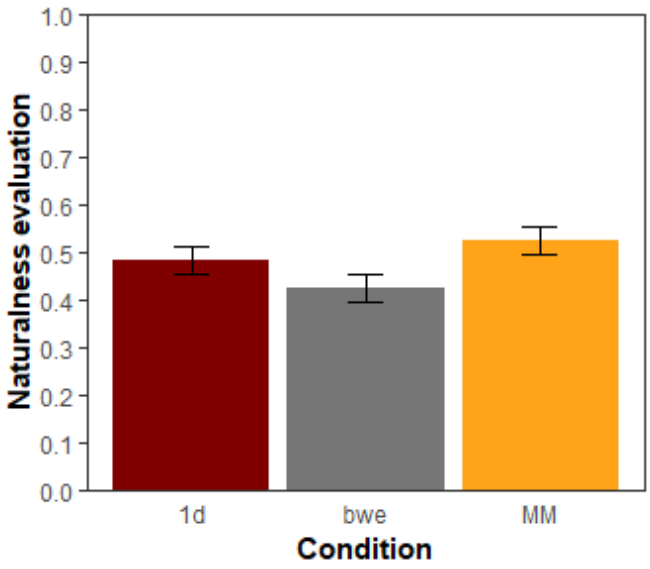

Figure 18 : Résultats de l'aspect naturel sur le degré de réalisme du modèle acoustique

2.1.2. Effet du genre de la voix de synthèse (QR 1)

Pour rappel, notre question de recherche était : « **Est-ce que les divers degrés de réalismes dans la génération des hautes fréquences impactent la perception de l'aspect naturel différemment pour les stimuli masculins VS pour les stimuli féminins ?** ». Un effet significatif a été mis en avant pour l'effet principal du genre de la voix de synthèse ($X^2 = 17.439$, $p < .001$). Les contrastes linéaires révèlent que la comparaison est significative. Comme illustré dans la *Figure 19*, les stimuli générés avec la voix de synthèse masculine sont considérés comme plus naturels que les stimuli générés avec la voix de synthèse féminines ($Z = - 21.860$; $p < .001$) **Nous constatons** que **l'aspect naturel des stimuli est impacté différemment selon qu'il s'agisse de stimuli masculins ou de stimuli féminins**.

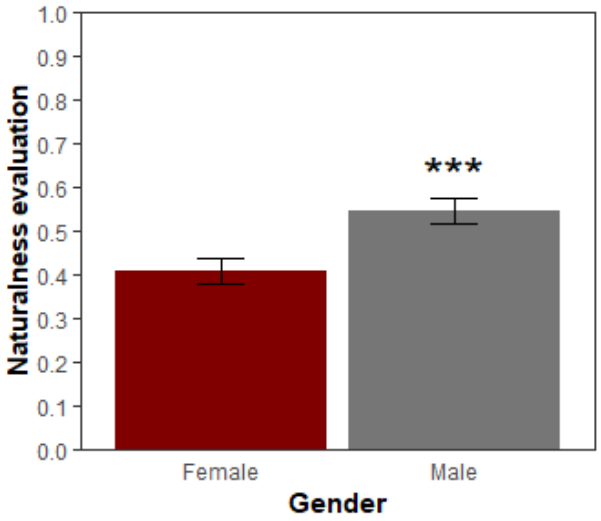

Figure 19 : Résultats de l'aspect naturel sur le genre de la voix de synthèse

2.1.3. Effet de la qualité vocale (QR 2)

Pour rappel, notre question de recherche était : « **Est-ce que les divers degrés de réalismes dans la génération des hautes fréquences impactent la perception de l'aspect naturel différemment pour les stimuli de voix pressées VS de voix modales ?** ». Aucun effet significatif n'est mis en avant pour l'effet principal de la qualité vocale ($X^2 = 0.725$, $p > .05$). Étant donné que l'effet n'est pas significatif, aucun contraste linéaire n'a été réalisé.

2.1.4. Fiabilité inter juge (QR 4)

Pour rappel, notre question de recherche était : « ***Les réponses des différents participants seront-elles équivalentes ?*** ». La fiabilité inter-juges a été testée avec le *likelihood ratio*. Ce test permet de comparer deux modèles. Dans notre cas, les deux modèles comparés sont ceux avec et sans effet random, en d'autres termes, avec ou sans participants. Un effet significatif du participant a été objectivé ($X^2 = 576.34$, $p < .001$). Nous observons que **les participants n'évaluent pas de la même manière les stimuli**.

2.1.5. Effets d'interaction

Nous avons investigué l'interaction entre le degré de réalisme du modèle acoustique et le type de phonème(section 2.1.5.1.), et l'interaction entre le type de phonème et le genre de la voix de synthèse (section 2.1.5.2.). Le tableau 6 montre les résultats obtenus.

Variable indépendantes	Khi Carré (X^2)	Degrés de liberté (Df)	Probabilité (> X^2)
Degré de réalisme du modèle acoustique * Type de phonème	24.581	8	**0.002****
Type de phonème * Genre de la voix de synthèse	14.412	4	**0.006****

Tableau 6 : Résultats des effets d'interaction

L'interaction entre le type de phonème est une analyse complémentaire pour laquelle nous n'avions pas émis d'hypothèse ou de question de recherche. Elle ne sera pas traitée dans cette partie « résultats », ni dans la partie « discussion ». Les résultats de cette analyse complémentaire se trouvent en *Annexe 6*.

2.1.5.1. Interaction entre le degré de réalisme du modèle acoustique et le phonème

Comme indiqué dans le tableau 6, l'interaction entre le degré de réalisme du modèle acoustique et le phonème est significative. Afin d'approfondir l'analyse, des contrastes linéaires entre les phonèmes ont été réalisés. Cela signifie qu'une comparaison deux à deux a été réalisée pour chacun des 5 phonèmes.

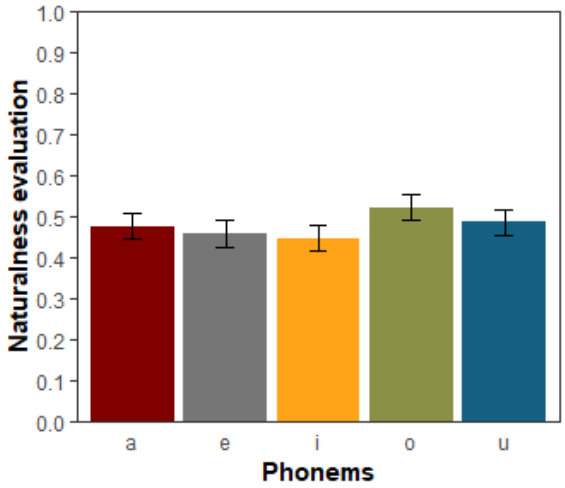

Figure 20 : Score moyen attribué pour l'aspect naturel de chaque phonème

La *Figure 20* permet de voir que le phonème considéré comme le plus naturel est le [o], tandis que le phonème considéré le moins naturel est le [i]. D'un point de vue statistique, les contrastes linéaires mettent en évidence plusieurs différences significatives. Plus précisément, le score d'aspect naturel du [a] n'est pas significativement différent de celui du [e] ($Z = 1.726$, $p > .05$), ni du [u] ($Z = - 1.156$, $p > .05$), mais il diffère significativement du [i] ($Z = 2.853$, $p < .05$) et du [o] ($Z = - 4.660$, $p < .001$). L'aspect naturel du [e] n'est pas significativement différent de celui du [i] ($Z = 1.116$, $p > .05$), mais il est significativement différent de celui du [o] ($Z = - 6.386$, $p < .001$) et du [u] ($Z = - 2.884$, $p < .05$). Concernant le [i], des différences significatives ont été relevées avec le [o] ($Z = - 7.537$, $p > .05$) et avec le [u] ($Z = - 4.020$, $p < .001$). Le score du [o], quant à lui, est significativement différent de celui du [u] ($Z = 3.514$, $p < .05$).

L'interaction a ensuite été étudiée entre le degré de réalisme du modèle acoustique et le type de phonème dans le but de répondre à notre question de recherche (QR 3). La *Figure 21* expose les scores obtenus pour l'aspect naturel de chaque phonème, selon les modèles 1D, 3D (MM) et BWE.

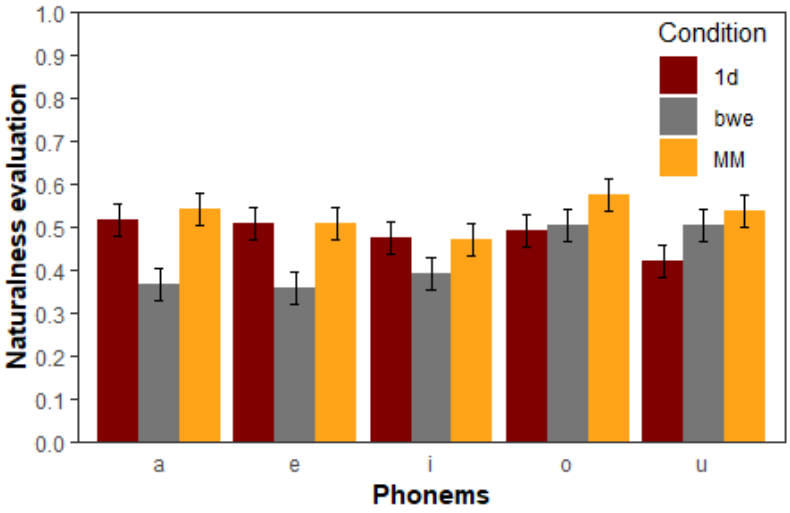

Figure 21 : Score moyen attribué pour l'aspect naturel de chaque phonème en fonction du degré de réalisme du modèle acoustique (1D, 3D, BWE)

Nous pouvons observer que le modèle 3D est considéré comme plus naturel, suivi du modèle 1D puis du modèle BWE pour les phonèmes [a], [e], [i]. En revanche, pour les phonèmes [o] et [u], le modèle 1D et le modèle BWE sont au même niveau pour le phonème [o] bien que le modèle 3D reste supérieur. Pour le phonème [u], le modèle 1D est moins naturel que le modèle BWE, qui lui-même est perçu comme moins naturel que le modèle 3D. L'interaction entre le degré de réalisme physique du modèle acoustique et le type de phonème étant significative, des contrastes linéaires ont été réalisés afin d'approfondir l'analyse.

Les résultats complets se trouvent en *Annexe 7*. Il existe plusieurs différences significatives. En ce qui concerne le phonème [a], il existe une différence significative entre le modèle 1D et le modèle BWE ($Z = 8.561$, $p < .001$) et entre le modèle 3D et le modèle BWE ($Z = - 9.773$, $p < .001$). Pour le phonème [e], des différences significatives ont été observées entre le modèle 1D et le modèle BWE ($Z = 8.397$, $p < .001$), ainsi qu'entre le modèle BWE et le modèle 3D ($Z = - 8.429$, $p < .001$). Le phonème [i] montre une différence significative entre le modèle 1D et le modèle BWE ($Z = 4.851$, $p < .001$), également entre le modèle BWE et le modèle 3D ($Z = - 4.584$, $p < .001$). À propos du phonème [o], une différence significative a été démontrée entre le modèle BWE et le modèle 3D ($Z = - 4.142$, $p < .05$), mais aussi entre le modèle 1D et le modèle 3D ($Z = - 4.867$, $p < .001$). Pour le phonème [u], des différences significatives ont été trouvées entre le modèle 1D et le modèle BWE ($Z = - 4.941$, $p < .001$) ainsi qu'entre le modèle 1D et le modèle 3D ($Z = - 6.666$, $p < .001$). **Nous remarquons** que **les divers degrés de réalismes dans la génération des hautes fréquences impactent la perception de l'aspect naturel différemment pour les phonèmes [a], [e], [i], [o], [u]** (QR 3).

VI. DISCUSSION

1. Rappel des objectifs de l'étude et de la méthodologie

Pour rappel, ce mémoire s'inscrit dans le cadre d'un projet de développement d'un outil de synthèse articulatoire à large bande dont l'aspect se veut le plus naturel possible. Son but est de mieux comprendre les HF (> 5 kHz), et de déterminer si l'aspect naturel de la parole varie en fonction du modèle acoustique exploité pour générer une parole de synthèse. Afin de répondre à cette problématique, les HF ont été inclues dans trois modèles acoustiques (1D, 3D, BWE). La partie expérimentale de ce mémoire visait à déterminer si la perception de l'aspect naturel des différents phonèmes était influencée par le modèle acoustique utilisés. Pour ce faire, deux tâches perceptives ont été réalisées par 40 jeunes adultes. Les tâches se basaient sur 20 stimuli synthétiques : 5 phonèmes ([a], [e], [i], [o], [u] x 2 genres (homme, femme) x 2 types de voix (modale, pressée) ; générés à l'aide de 3 modèles acoustiques (1D, 3D, BWE). La première tâche était une comparaison par paires, la seconde tâche était une évaluation de la naturalité de chaque stimulus. L'objectif de ces deux tâches était de déterminer si les HF jouaient effectivement un rôle dans la perception de l'aspect naturel, mais également de démontrer si un modèle, notamment le modèle acoustique 3D, offrait une parole de synthèse davantage naturelle que les deux autres modèles acoustiques 1D et BWE.

Trois hypothèses ont été émises concernant le modèle acoustique, et 4 questions de recherches ont été exprimées à propos du type de phonème, du type de voix, du genre de la voix de synthèse, et de la fiabilité inter-juges. Comme précédemment mentionné, les différentes questions de recherche n'ont pas toutes été testées dans nos deux expériences.

Cette partie « discussion » s'articule en deux parties. D'une part, nous synthétiserons et interpréterons les résultats obtenus pour chaque hypothèse et question de recherche. D'autre part, nous présenterons les limites et les biais méthodologiques rencontrés dans ce mémoire, dans le but de proposer des perspectives d'amélioration pour le futur.

2. Interprétation des résultats

2.1. Degré de réalisme physique du modèle acoustique (H1, H2, H3)

Nous avons souhaité investiguer l'influence des différents types de modèles acoustiques sur l'aspect naturel de la parole de synthèse. Des études semblables ont été réalisées. Le mémoire qui précède le nôtre (Hoornaert, 2021) a souhaité étudier les modélisations acoustiques 1D et 3D en réalisant deux tâches perceptives également, néanmoins, aucun effet significatif du modèle n'a pu être objectivé. L'étude de Gully (2017) a cherché à démontrer les différences en terme d'aspect naturel entre plusieurs degrés de réalismes physiques, notamment le modèle 3D, en réalisant des tâches de jugement à l'aide d'un curseur. Les résultats de cette étude (Gully, 2017) montrent que le modèle 3D est considéré comme étant le plus naturel. Nos résultats sont en accords avec les résultats de Gully (2017). Dans notre mémoire, le modèle 3D est jugé comme étant le plus naturel vis-à-vis des deux autres modèles acoustiques utilisés (1D et BWE).

Pour notre première expérience, des différences significatives ont été objectivées. Les résultats **confirment nos hypothèses H1, H2, et H3**. Le modèle 3D est jugé comme étant le plus naturel, suivi du modèle 1D, et du modèle BWE, ce dernier étant jugé comme le moins naturel. Bien que le résultat est significatif, il est important de noter que cette tendance n'est pas systématiquement observée pour tous les phonèmes. En effet, dans le point suivant (VI, 2.4), nous exposons les variations présentes entre le modèle 1D et le modèle 3D. En réalité, pour les phonèmes [o] et [u], le modèle 1D obtient une évaluation relativement basse en terme d'aspect naturel. C'est pourquoi, nous supposons que cela a influencé nos résultats concernant nos trois hypothèses à propos des degrés de réalisme physique des modèles acoustiques. Nous estimons que le modèle 1D est tellement faible en terme de naturalité pour certains phonèmes, que l'analyse générale du modèle 1D est tirée vers le bas, influencée par ces variations entre les phonèmes. Nous supposons que ce résultat général n'est pas observée pour tous les phonèmes, car elle est fortement influencée par les phonème [o] et [u].

Pour notre seconde expérience, des différences significatives ont également été révélées. Les résultats **confirment nos hypothèses H1, H2 et H3**. Le modèle 3D est considéré comme le plus naturel, suivi du modèle 1D, et du modèle BWE qui est lui jugé comme étant le moins

naturel. Le modèle 3D est le modèle qui est jugé le plus naturel, notamment car il s'agit du seul modèle qui, d'une part, tient compte de la géométrie précise du conduit vocal, permettant de produire des stimuli synthétiques proches de la réalité, et d'autre part, modélise correctement les HF qui jouent un rôle dans l'aspect naturel de la parole (Arnela et al., 2019 ; Freixes et al., 2018).

2.2. Genre de la voix de synthèse (QR 1)

Nous avons émis une question de recherche qui était : « **Est-ce que les divers degrés de réalismes dans la génération des hautes fréquences impactent la perception de l'aspect naturel différemment pour les stimuli masculins VS pour les stimuli féminins ?** ». Nous avons investigué cette question de recherche dans notre seconde tâche expérimentale. Les résultats montrent un effet significatif du genre de la voix de synthèse. En effet, la voix masculine est plus souvent considérée comme étant davantage naturelle que la voix féminine. Nous constatons que **l'aspect naturel des stimuli est impacté différemment selon qu'il s'agisse de stimuli masculins ou de stimuli féminins**.

Dans la « revue de la littérature » (II, chap. 2, 4.2), nous avons remarqué que la parole féminine contenait davantage de HF, notamment sur base de la physiologie humaine (Smith, 2005). Nos stimuli ont été générés avec VocalTractLab, logiciel permettant de contrôler de nombreux paramètres (Birkholz, 2013) dans le but de construire des stimuli proches de la réalité. Dans notre cas, nous pouvons supposer que la génération des stimuli féminins n'a pas été aussi bien modélisée que les stimuli masculins. Cela soulève la question de la construction des stimuli qui a peut-être été plus difficile pour les stimuli féminins que pour les stimuli masculins, offrant des stimuli moins naturels que ceux produits par le locuteur masculin. En effet, créer des stimuli synthétiques avec un réalisme physique est une tâche difficile, chronophage, et qui offre des stimuli parfois incomplets (Van Niekerk et al., 2022). Néanmoins, une différence significative entre les deux genres de voix de synthèse a été exposée dans nos résultats, et cela répond à notre question de recherche. Nous n'avions émis aucune hypothèse concernant le genre de la voix de synthèse qui était susceptible d'être considéré comme étant le plus naturel, car à ce jour, aucune étude n'a comparé des stimuli féminins et des stimuli masculins en se basant sur trois modèles acoustiques.

2.3. Qualité vocale (QR 2)

Nous avons souhaité déterminer si la qualité vocale jouait un rôle dans la perception de l'aspect naturel de la parole synthétique. Pour ce faire, nous avons émis comme question de recherche : « **Est-ce que les divers degrés de réalismes dans la génération des hautes fréquences impactent la perception de l'aspect naturel différemment pour les stimuli de voix pressées VS de voix modales ?** ». Aucun effet significatif n'a été découvert suite au modèle linéaire mixte, par conséquent, aucun contraste linéaire n'a été réalisé pour investiguer plus en profondeur**. Nous ne sommes pas en mesure d'observer si l'aspect naturel des stimuli est impacté différemment selon qu'il s'agisse de stimuli en voix modale ou de stimuli en voix pressée**.

Une étude de Kadiri et al., (2020), ne portant pas sur une parole de synthèse a révélé des résultats similaires. L'objectif de cette étude était de classifier les trois types de phonation (modal, pressé, soufflé) d'une part pour la voix parlée et d'autre part pour la voix chantée. Les résultats démontrent qu'il est difficile de comparer des types de qualité vocale car beaucoup de paramètres, tels que l'amplitude, les fréquences ou encore les quotients d'ouvertures, sont à prendre en compte pour y arriver. Les auteurs mettent d'ailleurs en avant l'importance de trouver des paramètres qui aideraient à percevoir les différences entre les types de voix pour faciliter ce classement. Cette étude nous laisse supposer qu'éventuellement nos participants se sont également trouvés face à une difficulté pour distinguer la voix modale de la voix pressée, surtout que les stimuli étaient des phonèmes isolés, ce qui représente un son très court, qui ne laisse peut-être pas le temps d'analyser en détail la qualité vocale.

2.4. Type de phonème (QR 3)

Sur base de l'étude qui nous précède (Hoornaert, 2021) où une différence significative avait été observée en ce qui concerne le type de phonème, notamment entre les voyelles, nous avons souhaité investiguer l'effet du type de phonème à notre tour. Nous avons émis la question de recherche suivante : « **Est-ce que les divers degrés de réalismes dans la génération des hautes fréquences impactent la perception de l'aspect naturel différemment pour les phonèmes [a], [e], [i], [o], [u] ?** ». En d'autres termes, nous souhaitions déterminer si d'une part l'aspect naturel était jugé différemment selon le phonème, mais également si d'autre part,

les phonèmes générés avec le modèle 3D par exemple, étaient jugés comme plus naturels que les phonèmes générés par le modèle 1D ou par le modèle BWE.

Pour notre première expérience, des différences significatives ont été démontrées entre les différents modèles acoustiques pour plusieurs phonèmes. Nous observons que **les divers degrés de réalismes dans la génération des hautes fréquences impactent la perception de l'aspect naturel différemment pour les phonèmes [a], [e], [i], [o], [u]**. En effet, les trois modèles acoustiques n'offrent pas la même naturalité pour tous les phonèmes. Il existe des différences entre eux, sauf pour les phonèmes [i] et [o] où deux modèles sont considérés comme étant identiques (1D-3D, 1D-BWE), mais aussi pour le phonème [e] où la différence entre les modèles 1D-3D est minime. Nous pouvons également affirmer qu'il existe des différences au niveau du classement. Cela signifie que même si aucune analyse statistique n'a été réalisée pour comparer les phonèmes entre eux, nous pouvons dire que le classement n'est pas le même selon les phonèmes. Le modèle 1D semble être le plus souvent en première position, ainsi que le modèle 3D pour d'autres phonèmes. Le modèle BWE quant à lui n'est jamais choisi comme étant le plus naturel. Cela signifie que l'aspect naturel des phonèmes est impacté selon les divers degrés de réalismes physiques des modèles acoustiques, ainsi que par le phonème évalué.

Pour notre seconde expérience, des différences significatives ont été objectivées. D'une part, il existe un effet significatif du type de phonème. En effet, quelles que soient les autres variables, les participants considèrent les phonèmes avec un degré de naturel différent. Sur base de ces résultats, **nous remarquons que les phonèmes impactent différemment l'évaluation de la naturalité**. Nous pouvons supposer que ces différences d'aspect naturel en fonction du phonème s'expliquent par la modélisation en voix de synthèse. Nous supposons par exemple, que la voyelle [i] est une voyelle où l'écart entre le formant et la fréquence fondamentale est relativement grand, cet écart peut rendre un aspect moins naturel lorsqu'il s'agit d'une modélisation synthétique. Yamasaki et al., (2017) ont réalisé une étude sur l'évaluation perceptive de l'aspect naturel de voyelles synthétisées. Les résultats de cette étude mettent en avant l'importance d'intégrer le *Jitter*[12] et le *Shimmer*[13] dans les voix de synthèse afin d'offrir une parole considérée comme étant plus naturelle. Nos stimuli ont été soumis à des variations de fréquence fondamentale, mais le *Jitter* et le *Shimmer* en tant que tels n'étaient pas utilisés

[12] Le *Jitter* correspond au degré de variabilité de la fréquence. Il permet de mesurer la raucité.

[13] Le *Shimmer* correspond au degré de variabilité de la pression sonore. Il permet de mesurer le souffle.

(Rémi Blandin, communication personnelle, le 25 mai 2022). Nous estimons qu'une amélioration de l'implémentation de ces paramètres dans la création des stimuli est un enjeu pour une éventuelle future étude basée sur ce sujet. Bien que l'étude de Yamasaki et al., (2017) ne s'est basée que sur une seule voyelle, nous pouvons supposer qu'améliorer des paramètres comme le *Jitter* ou le *Shimmer* permettrait de limiter ces différences d'aspect naturel entre les phonèmes. D'autre part, un effet significatif du degré de réalisme a été objectivé sur certains phonèmes. Les modèles acoustiques n'ont pas les mêmes effets suivant les différents phonèmes. De façon générale, le modèle 3D est toujours considéré comme étant le modèle qui génère des stimuli synthétiques les plus naturels, cependant, l'aspect naturel des stimuli générés par les modèles 1D et BWE varient. Nous notons que **les divers degrés de réalismes dans la génération des hautes fréquences impactent la perception de l'aspect naturel différemment pour les phonèmes [a], [e], [i], [o], [u]**. Nos résultats montrent que le modèle 3D est plus naturel peu importe le phonème, cela confirme qu'il s'agit du modèle permettant de décrire les propriétés acoustiques du conduit vocal le plus proche de la réalité (Birkholz, 2013). Nous supposons que la variation entre le modèle 1D et le modèle BWE s'explique par leur manque de complexité ou leur méthode approximative qui offre des résultats moins stables. La création de stimuli synthétiques étant compliquée à réaliser, avec des résultats parfois insatisfaisants (Van Niekerk et al., 2022), nous pouvons supposer que les modèles 1D et BWE soient plus sensibles à ces variations de résultats, satisfaisants ou non.

2.5. Fiabilité inter-juges (QR 4)

Nous avons souhaité investiguer la fiabilité inter-juges pour examiner la présence ou non d'homogénéité entre les réponses données par l'ensemble des participants. Pour rappel, nous avions posé la question de recherche suivante : « ***Les réponses des différents participants seront-elles équivalentes ?*** ». Les résultats de notre seconde expérience démontrent que **les participants n'ont pas évalué de la même manière les différents stimuli**. Nous pouvons supposer que cela est dû à notre large échantillon, avec des participants âgés de 19 à 24 ans, leur sensibilité aux HF n'est donc pas la même. Les résultats de l'audiométrie tonale (*Annexe 4*) ont révélé que les participants âgés de 19 ans disposaient d'une meilleure sensibilité auditive que les participants âgés de 22 ou 23 ans par exemple. Nous pouvons également supposer que la concentration durant la tâche perceptive a varié entre les différents participants. De plus, nous pouvons supposer que tous les participants ne possédaient pas la même référence interne de ce qu'était une parole naturelle.

3. Limites et perspectives

Plusieurs limites sont à mentionner dans ce mémoire. Les résultats présentés dans la section précédente (V, 1 et 2) sont à prendre avec précaution, notre étude se voulait exploratoire. Différentes perspectives sont suggérées afin d'améliorer le plan expérimental, en cas d'études ultérieures portant sur le même sujet.

3.1. Absence de certaines analyses statistiques

Dans les partie « résultat » (V) et « discussion » (VI), nous n'avons pas pu tester toutes nos hypothèses pour l'expérience 1. Pour rappel, cette tâche était une comparaison par paires. Ce manque d'analyses statistiques est une limite du mémoire. En effet, toutes les comparaisons statistiques n'ont pas été possibles. Le plan expérimental était limité afin que l'expérience soit humainement possible, par conséquent, nous avons dû réaliser des statistiques avec un modèle relativement complexe à manipuler, qui ne permettait pas de répondre à toutes nos questions de recherche sur base du plan expérimental. À titre d'exemple, l'effet du genre de la voix de synthèse n'a pas pu être évalué. Cela explique pourquoi nous n'avons aucune information à ce niveau-là pour la première tâche expérimentale. L'encodage de ce modèle étant assez lourd, nous n'avons pu tester que nos hypothèses sur les degrés de réalismes physiques et notre question de recherche sur le type de phonème, bien que cette analyse sur le type de phonème a, elle aussi, été limitée à comparer les phonèmes indépendamment les uns des autres.

3.2. Modalité de passation des tâches

Une perspective serait de limiter le nombre de stimuli. La première tâche était composée de 120 paires de stimuli, en d'autres termes, le participant entendait au minimum 240 stimuli s'il n'écoutait qu'une seule fois chaque stimuli. La seconde tâche était composée de 120 stimuli. Au total, 360 stimuli au minimum ont été entendus par chaque participant. Au niveau cognitif, ce sont des tâches très complexes, mais aussi différentes. La comparaison de deux stimuli ne fait pas intervenir le même niveau de cognitif qu'attribuer une note à un stimuli. Il nous semble que le nombre de stimuli devrait être réduit pour avoir un testing plus court, et peut-être plus fiable. Une alternative serait de mesurer le niveau de concentration des participants, afin de s'assurer que leur concentration est la même en début et en fin de testing. Des participants nous

ont avoué avoir été démoralisés face à la seconde tâche, car après 120 comparaisons, ils désiraient terminer le testing rapidement. C'est pourquoi, limiter le nombre de stimuli nous parait une suggestion intéressante pour des études ultérieures.

3.3. Environnement de l'étude

La cabine audiométrique disposait d'un chauffage fixé au mur. Celui-ci diffusait un bruit continu de façon aléatoire. Il est survenu plusieurs fois durant l'audiométrie ou durant les deux tâches expérimentales. Puisque ce bruit survenait de façon aléatoire, il nous était impossible de faire quoique ce soit pour l'éviter durant nos testings. Nous avons suivi les conseils de Xavier Kaiser en éteignant complètement le chauffage avant de débuter nos testings afin d'éviter que ces bruits surviennent. Malheureusement, les précautions se sont montrées insuffisantes. Ce biais lié à l'environnement de l'étude n'a eu aucun effet sur la récolte de nos données. Le bruit de fond était de 25 dB SPL dans la cabine audiométrique, ce qui rend nos données interprétables.

VII. CONCLUSION

Ce mémoire s'est inscrit dans le cadre d'un projet de recherche qui avait pour but de développer un outil de synthèse articulatoire à large bande dont l'aspect serait le plus naturel possible, afin de mieux comprendre le lien entre la perception de la parole et les hautes fréquences (HF). Pour ce faire, les phonèmes ont été synthétisés avec trois modélisations physiques différentes : le modèle acoustique unidimensionnel (1D), le modèle acoustique tridimensionnel (3D) et le modèle d'algorithme d'extension (BWE). Le modèle 3D est le modèle qui semble le plus prometteur pour générer des stimuli synthétiques proches de la réalité grâce à ses différentes innovations qui sont possibles grâce à l'IRM. Il permet de représenter la gamme des HF (> 5 kHz) avec son réalisme acoustique. Dans ce mémoire, nous avons souhaité approfondir la question du rôle des HF dans la perception de l'aspect naturel de la parole, au travers de parole de synthèse.

Le premier objectif était de comparer l'aspect naturel de la parole obtenu avec les différentes modélisations physiques, dans un paradigme de comparaison par paires, où les stimuli générés avec chacun des modèles étaient comparés deux à deux. Un effet significatif du degré de réalisme physique du modèle acoustique a été mis en avant. Le modèle 3D a été considéré comme offrant les stimuli les plus naturels, suivi du modèle 1D, et le modèle BWE a été considéré comme offrant les stimuli les moins naturels. Bien que ces résultats nous permettent de confirmer nos hypothèses selon lesquelles les stimuli générés par le modèle 3D sont perçus comme plus naturels que les stimuli générés par le modèle 1D ; les stimuli générés par le modèle 3D sont perçus comme plus naturels que les stimuli générés par le modèle BWE ; et les stimuli générés par le modèle 1D sont perçus comme plus naturels que les stimuli générés par le BWE. Un effet significatif du type de phonème a également été objectivé, permettant de répondre à notre question de recherche qui mentionne que les divers degrés de réalismes physiques dans la génération des HF impactent la perception de l'aspect naturel différemment pour les phonèmes [a], [e], [i], [o], [u]. Nos résultats ont démontré des différences significatives entre les modèles acoustiques, sauf pour les phonèmes [i] et [o] où deux modèles sont considérés comme étant identiques (1D-3D, 1D-BWE), mais aussi pour le phonème [e] où la différence entre les modèles 1D-3D est légère.

Le second objectif avait pour but d'évaluer l'aspect naturel de chacun des stimuli générés avec chacun des trois modèles, à l'aide d'une échelle métrique allant de 0 (pas du tout naturel) à 100 (totalement naturel). Un effet significatif du degré de réalisme physique du modèle acoustique a été mis en avant : le modèle 3D est considéré comme étant le plus naturel, suivi du modèle 1D, lui-même suivi du modèle BWE. Un effet significatif du type de phonème a été objectivé : les phonèmes n'étaient pas évalués de la même façon en termes de naturalité. Une interaction a été démontrée entre le degré de réalisme physique du modèle acoustique et le type de phonème car les conditions n'ont pas les mêmes effets suivant les différents phonèmes. Un effet significatif du genre de la voix de synthèse a été mis en lumière, en effet les voix masculines étaient considérées comme étant plus naturelles que les voix féminines, et ce peu importe les différentes variables.

Précédemment, une étude (Gully, 2017), avait souhaité examiner l'effet de différents degrés de réalismes physiques des modèles acoustiques sur la perception de l'aspect naturel de la parole auprès d'auditeurs réels. Notre mémoire a permis de mettre en évidence la sensibilité auditive de participants âgés de 19 à 24 ans, mais a également permis de démontrer l'apport du modèle 3D dans l'aspect naturel des 5 phonèmes qui ont été évalués. Nos résultats sont en accords avec cette étude (Gully, 2017), qui avait également objectivé un aspect davantage naturel avec le modèle 3D, pourtant différent du nôtre. Néanmoins, nos résultats restent à considérer avec prudence au vu des différentes limites et des différents biais méthodologiques présents.

Ce mémoire a cherché à explorer le rôle des HF dans la perception des phonèmes selon différents degrés de réalismes physiques de modèles acoustiques.

VIII. BIBLIOGRAPHIE

- Alexander, J.M., Kopun, J.G., & Stelmachowicz, P.G., (2014). Effects of Frequency Compression and Frequency Transposition on Fricative and Affricate Perception in Listeners with Normal Hearing and Mild to Moderate Hearing Loss, *Ear Hear*, 35, pp.519–532, DOI:10.1097/AUD.0000000000000040.
- Antoine R., (2019), *Recherche d'une méthodologie d'analyse d'un lieu par et pour son ambiance sonore* (Mémoire Université de Liège, Gembloux Agro-Bio Tech), http://hdl.handle.net/2268.2/8308
- Apoux, F., & Bacon, S.P. (2004). Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise, *Journal of the Acoustical Society of America*, 116, pp.1671–1680, DOI:10.1121/1.1781329.
- Arnela, M., Dabbaghchian, S., Guasch, O., & Engwall, O. (2019). MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(12), pp.2173-2182. https://doi.org/10.1109/TASLP.2019.2942439.
- Arnela, M., Dabbaghchian, S., Guasch, O., & Engwall, O. (2019). MRI-based vocal tract representations for the three-dimensional finite element synthesis of diphthongs. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *27*(12), 2173-2182. https://doi.org/10.1109/TASLP.2019.2942439
- Bachhav, P., Todisco, M. & Evans, N. (2018a). Efficient Super-Wide Bandwidth Extension Using Linear Prediction Based Analysis-Synthesis, *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5429-5433, doi: 10.1109/ICASSP.2018.8462547.
- Bachhav, P., Todisco, M., Evans, N. (2018b) Artificial Bandwidth Extension with Memory Inclusion Using Semi-supervised Stacked Auto-encoders, *Proc. Interspeech*, pp.1185-1189, DOI: 10.21437/Interspeech.2018-2213.
- Badri, R., Siegel, J.H., & Wright, B.A. (2011). Auditory filter shapes and high-frequency hearing in adults who have impaired speech in noise performance despite clinically normal audiograms, *Journal of the Acoustical Society of America*, 129 (2), pp.852–863, DOI:10.1121/1.3523476.
- Bagot, J-D. (1999). Information, sensation et perception, *Armand Colin*, Paris, 192 pages.
- Baken, R. J., & Orlikoff, R. F. (2000). Clinical measurement of speech and voice (2nd ed.). *Singular Thomson Learning*, Boston, 604 pages.
- Baumann, O., & Belin, P. (2008). Perceptual scaling of voice identity: common dimensions for different vowels and speakers. *Psychological research*, *74*(1), pp.110–120. https://doi.org/10.1007/s00426-008-0185-z.
- Belin, P., Bestelmeyer, P.E., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British journal of psychology (London, England: 1953)*, *102*(4), pp.711–725. https://doi.org/10.1111/j.2044-8295.2011.02041.x.

- Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: neural correlates of voice perception, *Trends in Cognitive Science*, 8, pp.129–135, DOI: 10.1016/j.tics.2004.01.008
- Bellinghausen, C., Fangmeier, T., Schröder, B., Keller, J., Drechsel, S., Birkholz, P., Tebartz van Elst, L., Riedel, A., & Nussbruch, T. (2019). On the Role of Disfluent Speech for Uncertainty in Articulatory Speech Synthesis, *Proceedings of DiSS 2019*, pp.39-42, DOI :10.21862/diss-09-011-bell-etal.
- Ben Aissa, B. (2020), Classification des domaines protéiques par techniques d'apprentissage profond, *Faculté des Sciences Exactes et des Sciences de la Nature et de la Vie (FSESNV)*, Université Mohamed Khider – BISKRA, 73 pages.
- Berlin, C.I. (1982). Ultra-audiometric hearing in the hearing impaired and the use of upward-shifting translating hearing aids, *The Volta Revue,* 84, pp.352–353.
- Best, V., Carlile, S., Jin, C., & Van Schaik, A. (2005). The role of high frequencies in speech localization. *The Journal of the Acoustical Society of America*, 118(1), pp.353-363. https://doi.org/10.1121/1.1926107.
- Birkholz, P. (2007). Control of an articulatory speech synthesizer based on dynamic approximation of spatial articulatory targets. *Interspeech 2007 - Eurospeech*, pp.2865–2868, https://doi.org/10.1007/978-3-540-76442-7_16.
- Birkholz, P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis, *PlosOne*, 8 (4), pp.1-17, https://doi.org/10.1371/journal. pone.0060603.
- Birkholz, P., (2017). Vocal Tract Lab (version 2.2). Disponible sur http://www.vocaltractlab.de/ (page consultée le 27/04/2021).
- Birkholz, P., & Drechsel, S. (2021). Effects of the piriform fossae, transvelar acoustic coupling, and laryngeal wall vibration on the naturalness of articulatory speech synthesis. *Speech Communication*, *132*, 96-105. https://doi.org/10.1016/j.specom.2021.06.002
- Birkholz, P., Jackèl, D. (2004). Influence of temporal discretization schemes on formant frequencies and bandwidths in time domain simulations of the vocal tract system. *Interspeech 2004-ICSLP*, pp.1125–1128.
- Birkholz, P., Jackèl, D., and Kröger, B.J. (2006). Construction and control of a three-dimensional vocal tract model. *International Conference on Acoustics, Speech, and Signal Processin*g, pp.873–876, DOI:10.1109/ICASSP.2006.1660160.
- Birkholz, P., Kröger, B.J., Neuschaefer-Rube, C. (2011b). Model-based reproduction of articulatory trajectories for consonant-vowel sequences. *IEEE Transaction on Audio Speech Lang. Process.,* 19, pp.1422–1433, DOI:10.1109/TASL.2010.2091632.
- Birkholz, P., Kröger, B.J., Neuschaefer-Rube, C., (2011a). Articulatory synthesis of words in six voice qualities using a modified two-mass model of the vocal folds, *First International Workshop on Performative Speech and Singing Synthesis*.
- Birkholz, P., Kröger, B.J., Neuschaefer-Rube, C., (2011c). Synthesis of breathy, normal, and pressed phonation using a two-mass model with a tri-angular glottis*, Interspeech-2011*, pp.2681–2684.
- Birkholz, P., Martin, L., Willmes, K., Kröger, B.J., Neuschaefer-Rube, C., (2015). The contribution of phonation type to the perception of vocal emotions in German: an

articulatory synthesis study, *The Journal of the Acoustical Society of America,* 137, pp.1503–1512, https://doi.org/10.1121/1.4906836.

- Birkholz, P., Martin, L., Xu, Y., Scherbaum, S., & Neuschaefer-Rube, C. (2017). Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis, *Computer Speech & Language*, *41*, pp. 116-127, https://doi.org/10.1016/j.csl.2016.06.004.
- Boyd-Pratt, H. A., & Donai, J. J. (2020). The perception and use of high-frequency speech energy: Clinical and research implications. *Perspectives of the ASHA Special Interest Groups*, *5*(5), 1347-1355.
- Bricker, P.D., & Pruzansky, S. (1976). Speaker recognition, *Contemporary issues in experimental phonetics*, New York, pp.295–326.
- Briot J-P, (2019), Apprentissage profond et génération musicale, *Hors-série Intelligence artificielle*, pp.30-37, hal-02267790v2.
- Briot J-P. (2019). Apprentissage profond et génération musicale, *Hors-série Intelligence artificielle*, pp.30-37.
- Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*, pp.305–327, DOI: 10.1111/j.2044-8295.1986.tb02199.x.
- Burton, A.M., Bruce, V., & Johnston, R.A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, *81*, pp.361–380, DOI: 10.1111 /j.2044-8295.1990.tb02367.x
- Burton, A.M., Jenkins, R., & Schweinberger, S.R. (2011). Mental representations of familiar faces. *British Journal of Psychology*, *102*, pp.943–958. doi:10.1111/j.2044-8295.2011.02039.x.
- Castellengo, M. (2015). Ecoute musicale et acoustique : Avec 420 sons et leurs audiogrammes décryptés. *Eyrolles*.
- Delvaux, V., & Pillot-Loiseau, C. (2020). Perceptual Judgment of Voice Quality in Nondysphonic French Speakers: Effect of Task-, Speaker- and Listener-Related Variables. *Journal of voice: official journal of the Voice Foundation*, *34*(5), 682–693. https://doi.org/10.1016/j.jvoice.2019.02.013.
- Dunn, H.K., White, S.D. (1940). Statistical measurements on conversational speech, Journal *of the Acoustical Society of America*, 11, pp.278–283, https://doi.org/10.1121/1.1916034.
- Ethofer, T., Van De Ville, D., Scherer, K., & Vuilleumier, P. (2009). Decoding of emotional information in voice-sensitive cortices, *Current Biology*, *19*, pp. 1028–1033. DOI: 10.1016/j.cub.2009.04.054
- Ferrieux C. (2011), Transformation de la voix humaine, https://interstices.info/transformation-de-lavoix-humaine
- Fletcher, H., & Galt, R. H. (1950). The perception of speech and its relation to telephony*, Journal of the Acoustical Society of America, 22, pp. 89-151,* https://doi.org/10.1121 /1.1906605
- Fletcher, H., & Steinberg, J.C. (1930). Articulation testing methods, *The Journal of the Acoustical Society of America*, 1(17), pp.1-48, https://doi.org/10.1121/1.1915183.

- Fowler, C.A., Saltzman, E., (1993), Coordination and coarticulation in speech production, *Language and Speech*, 36, pp.171–195, https://doi.org/10.1177/002383099303600304.
- Freixes, M., Alías, F. & Socoró, J.C. (2021). Contribution of vocal tract and glottal source spectral cues in the generation of happy and aggressive (a) vowels*, IberSPEECH 2021*, pp.240-244, doi: 10.21437/IberSPEECH.2021-51.
- Freixes, M., Arnela, M., Socoró, J. C., Alías, F., & Guasch, O. (2019). Glottal Source Contribution to Higher Order Modes in the Finite Element Synthesis of Vowels. *Applied Sciences*, *9*(21), 4535. https://doi.org/10.3390/app9214535
- Freixes, M., Arnela, M., Socoró, J. C., Pujol, F. A., & Guasch, O. (2018). *Influence of tense, modal and lax phonation on the three-dimensional finite element synthesis of vowel /a/* [Paper presentation]. IberSPEECH 2018, Barcelona, Spain.
- Geiser, B. (2012). Paths toward HD-voice communication, *International Workshop on Acoustic Signal Enhancement (IWAENC),* pp.1–4.
- Gervais, H., Belin, P., Boddaert, N., Leboyer, M., Coez, A., Barthelemy, C., *et al*. (2004). Abnormal voice processing in autism: A fMRI study, *Nature Neuroscience*, *7*, pp. 801–802. DOI:10.1038/nn1291.
- Gordon, M. & Ladefoged, P. (2001). Phonation types: a cross-linguistic overview*, Journal of Phonetics*, 29 (4), pp.383-406, https://doi.org/10.1006/jpho.2001.0147.
- Grandjean, D., Sander, D., Pourtois, G., Schwartz, S., Seghier, M.L., Scherer, K.R., *et al*. (2005). The voices of wrath: Brain responses to angry prosody in meaningless speech, *Nature Neuroscience*, *8*, pp.145–146. https://doi.org/10.1038/nn1392.
- Green, D.M., Kidd, G. Jr., & Stevens, K.N. (1987). High-frequency audiometric assessment of a young adult population, *Journal of the Acoustical Society of America*, 81, pp.485–494. https://doi.org/10.1121/1.394914.
- Grossman, T., Oberecker, R., Koch, S. P., & Friederici, A. D. (2010). The developmental origins of voice processing in the human brain, *Neuron*, *65*, pp. 852–858. doi: 10.1016/j.neuron.2010.03.001.
- Gully A. J., (2017), Diphthong Synthesis using the Three-Dimensional Dynamic Digital Waveguide Mesh, PhD thèse, University of York. https://etheses.whiterose.ac.uk/20043/
- Hanson, H. (1997). Glottal characteristics of female speakers: Acoustic correlates. *The Journal of the Acoustical Society of America*, 101, pp.466–481. https://doi.org/10.1121/1.417991.
- Hauser, M. D. (1996). *The Evolution of Communication*, MIT press, Cambridge, 720p.
- Hecker, M.H. L. (1971). Speaker recognition: An interpretive survey of the literature, *ASHA Monographs*, 16, pp.1-103.
- Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception, *Trends in cognitive sciences*, 4 (4), pp. 131-138. https://doi.org/10.1016/S1364-6613(00)01463-7.
- Huang, J., (2001), Articulatory speech synthesis and speech production modelling [Doctoral dissertation, University of Illinois], ResearchGate,

https://www.researchgate.net/publication/234531171_Articulatory_speech_synthesis_and_speech_production_modelling

- Hunt, A.J., Black, A.W., (1996). Unit selection in a concatenative speech synthesis system using a large speech database. *IEEE International Conference on Acoustics, Speech, and Signal Processing*, pp.373–376, doi: 10.1109/ICASSP.1996.541110.
- Hunter, L.L., Monson, B.B., Moore, D.R., Dhar, S., Wright, B.A., Munro, K.J., Zadeh, L.M., Blankenship, C.M., Stiepan, S.M., & Siegel, J.H. (2020). Extended high frequency hearing and speech perception implications in adults and children. *Hearing research*, 397, https://doi.org/10.1016/j.heares.2020.107922.
- Imbery, C., Franz, S., Van de Par, S. & Bitzer, J. (2019) Auditory Facing Angle Perception: The Effect of Different Source Positions in a Real and an Anechoic Environment, *Acta Acustica united with Acustica*, 105 (3), pp.492-505, **DOI:** https://doi.org/10.3813/AAA.919331.
- Ioannidis L., Rouas J-L., & Desainte-Catherine M.. (2014). Caractérisation et classification automatique des modes phonatoires en voix chantée, *XXXèmes Journées d'études sur la parole*.
- Johnson, K. (2003). Acoustic & auditory phonetics (2nd ed.). *Blackwell Publishing*, Oxford, 192p., DOI: 10.1159/000078663.
- Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives, *Journal of the Acoustical Society of America*, 108, pp.1252–1263, DOI:10.1121/1.1288413.
- Kacha, A., Grenez, F., Schoentgen, J. (2005) Voice quality assessment by means of comparative judgments of speech tokens. *Proceedings of the European conference on speech communication and technology, Interspeech*, Lisbon (Portugal), 80, pp. 1733–1736.
- Kadiri, S. R., Alku, P., & Yegnanarayana, B. (2020). Analysis and classification of phonation types in speech and singing voice. *Speech Communication*, *118*, 33-47.
- Kreiman, J. (2011) Foundations of voice studies « An interdisciplinarity approach to voice production and perception », *Wiley–Blackwell*, 516p.
- Kreiman, J., Gerratt, B.R., & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech, Language, and Hearing Research*, *33*(1), pp.103-115. https://doi.org/10.1044/jshr.3301.103.
- Kreiman, J., Gerratt, B.R., Precoda, K., & Berke, G. S. (1992). Individual differences in voice quality perception, *Journal of Speech and Hearing Research,* 35, pp.512–520, DOI: 10.1044 /jshr.3503.512.
- Kröger, B. (1992). Minimal rules for articulatory speech synthesis. *Proceedings of EUSIPCO92, 1*, pp.331-334.
- Lamesch S. (2006). Caractérisation de la voix mixte en termes de mécanismes laryngés (Projet de fin d'étude), Université Pierre & Marie Curie, http://www.atiam.ircam.fr/Archives/Stages0506/SylvainLamesch.pdf
- Lasarcyk, E., Wollermann, C., Schröder, B. & Schade, U. (2013). On the Modelling of Prosodic Cues in Synthetic Speech – Effects on Perceived Uncertainty and Naturalness

?, *Conference: 10th International Workshop on Natural Language Processing and Cognitive Science*.

- Lei Z., Kennedy E., Fasanella L., Li-Jessen NY-K, & Mongeau L., (2019), Discrimination between Modal, Breathy and Pressed Voice for Single Vowels Using Neck-Surface Vibration Signals, *Applied Sciences*, 9(7), https://doi.org/10.3390/app9071505.
- Linden, D.E., Thornton, K., Kuswanto, C.N., Johnston, S.J., van de Ven, V., & Jackson, M.C. (2011). The brain's voices: Comparing nonclinical auditory hallucinations and imagery, *Cerebral Cortex*, *31*, pp.330–337. DOI: 10.1093/cercor/bhq097.
- Lippmann, R.P. (1996). Accurate consonant perception without mid-frequency speech energy, *IEEE Trans Speech Audio Proc 4*, pp.66–69. https://doi.org/10.1109/TSA.1996.481454.
- Maci, L. & Carusi, A. (2020). La fiabilité de l'audiométrie tonale liminaire.
- Manley, G. A. (2016). Comparative auditory neuroscience: Understanding the evolution and function of ears, *Journal of the Association for Research in Otolaryngology*, 18, pp.1–24. https://doi.org/10.1007/s10162-016-0579-3.
- Mesgarani, N., & Chang E.F., (2012). Selective cortical representation of attended speaker in multi-talker speech perception. *Nature*, 485, pp.233–236. https://doi.org/10.1038/nature 11020.
- Meunier C., (2007), Phonétique acoustique : Phonétique acoustique, Les dysarthries, pp.164-173, hal-00250272.
- Meunier, C. (2007). Phonétique acoustique : Phonétique acoustique, *Les dysarthries*, pp.164-173.
- Monson, B. B., Hunter, E. J., & Story, B. H. (2012b). Horizontal directivity of low-and high-frequency energy in speech and singing. *The Journal of the Acoustical Society of America*, 132(1), pp.433-441. https://doi.org/10.1121/1.4725963.
- Monson, B. B., Lotto, A. J., & Story, B. H. (2014). Gender and vocal production mode discrimination using the high frequencies for speech and singing, *Frontiers in psychology*, 5(1239), pp.1-7. https://doi.org/10.3389/fpsyg.2014.01239.
- Monson, B.B. (2011). High-frequency energy in singing and speech, University of Arizona, Arizona, 161p., http://hdl.handle.net/10150/202695.
- Monson, B.B., & Caravello, J. (2019a). The maximum audible low-pass cutoff frequency for speech. *The Journal of the Acoustical Society of America*, 146(6), pp.496-501. https://doi.org/10.1121/1.5140032.
- Monson, B.B., Hunter, E.J., Lotto, A.J., & Story, B.H. (2014). The perceptual significance of high-frequency energy in the human voice. *Frontiers in psychology*, 5(587), pp.1-11. https://doi.org/10.3389/fpsyg.2014.00587.
- Monson, B.B., Lotto, A.J., & Ternström, S. (2011). Detection of high-frequency energy changes in sustained vowels produced by singers. *The Journal of the Acoustical Society of America*, 129(4), pp.2263-2268. https://doi.org/10.1121/1.3557033.
- Monson, B.B., Lotto, A.J., & Story, B.H. (2012a). Analysis of high-frequency energy in long-term average spectra of singing, speech, and voiceless fricatives, *Journal of the Acoustical Society of America*, 132, pp.1754–1764, DOI :10.1121/1.4742724.

- Monson, B.B., Rock, J., Schulz, A., Hoffman, E., & Buss, E. (2019b). Ecological cocktail party listening reveals the utility of extended high-frequency hearing. *Hearing research*, (381), DOI:10.1016/j.heares.2019.107773.
- Moore, B.C., Fullgrabe, C., & Stone, M.A. (2010). Effect of spatial separation, extended bandwidth, and compression speed on intelligibility in a competing-speech task, *Journal of the Acoustical Society of America*, 128, pp.360–371, https://doi.org/10.1121/1.3436533.
- Moore, B.C.J. (2003). An introduction to the psychology of hearing, *Academic Press*, Amsterdam.
- Moore, B.C.J., & Tan, C.-T. (2003). Perceived naturalness of spectrally distorted speech and music, *Journal of the Acoustical Society of America*, 114, pp.408–419, DOI:https://doi.org /10.1121/1.1577552.
- Moore, B.C.J., Stone, M.A., Fullgrabe, C., Glasberg, B.R., & Puria, S. (2008). Spectro-temporal characteristics of speech at high frequencies, and the potential for restoration of audibility to people with mild-to-moderate hearing loss, *Ear Hear,* 29 (6), pp.907–922, DOI:10.1097/ AUD.0b013e31818246f6
- Moore, D.R., Hunter, L.L., & Munro, K.J. (2017). Benefits of Extended High-Frequency Audiometry for Everyone, *Hearing Journal*, 70 (3), pp.50-55. DOI: 10.1097/01.HJ.00 00513797.74922.42.
- Morsomme, D. (2021). Évaluation et prise en charge des troubles de la voix parlée LOGO0019-1 (Cours), Université de Liège.
- Murphy A., Yanushevskaya I., Chasaide A., & Gobl C., (2019). The Role of Voice Quality in the Perception of Prominence in Synthetic Speech, *Interspeech*, DOI:10.21437/Interspeech. 2019-2761.
- Neuhoff, J.G. (2003). Twist and Shout: Audible Facing Angles and Dynamic Rotation, *Ecol. Psychol*, 15, pp.335–351, DOI:10.1207/s15326969eco1504_7.
- Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar voices. *The Journal of the Acoustical Society of America*, *85*(2), pp.913-925. https://doi.org/10.1121/1.3 97564.
- Petkov, C.I., Kayser, C., Steudel, T., Whittingstall, K., Augath, M., & Logothetis, N.K. (2008). A voice region in the monkey brain, *Nature Neuroscience*, *11*, pp. 367–374. https://doi.org/10.1038/nn2043.
- Pittman, A.L. (2008). Short-term word-learning rate in children with normal hearing and children with hearing loss in limited and extended high-frequency bandwidths, *Journal of Speech, Language, and Hearing Research*, 51, pp.785–797, doi:10.1044/1092-4388(2008/056).
- Pulakka, H., Laaksonen, L., Yrttiaho, S., Myllyla, V., & Alku, P. (2012). Conversational quality evaluation of artificial bandwidth extension of telephone speech, *Journal of the Acoustical Society of America*, 132 (2), pp.848–861, DOI:10.1121/1.4730882.
- Rebillard G., (2021), Fonctionnement de la cochlée, http://www.cochlea.eu/cochlee/ fonctionnement.
- Recommandation biap 02/1 bis : Classification audiométrique des déficiences auditives disponible sur https://www.biap.org/fr/recommandations/recommandations/ct-02-

classification-des-deficiences-auditives/149-rec-02-01-fr-classification-audiometrique-des-deficiences-auditives/file

- Remacle, A. (2013), La charge vocale : De sa quantification à l'étude de son impact sur la fonction phonatoire et sur la qualité vocale, *Thèse, Faculté de Psychologie et des Sciences de l'Éducation Unité Logopédie de la Voix*, 377 pages.
- Remacle, A., Schoentgen, J., Finck, C., Bodson, A., & Morsomme, D. (2014) Impact of vocal load on breathiness: Perceptual evaluation, *Logopedics Phoniatrics Vocology*, 39 (3), pp.139-146, DOI: 10.3109/14015439.2014.884161.
- Remez, R.E., Rubin, P.E., Pisoni, D.B., & Carrell, T.D. (1981). Speech perception without traditional speech cues, *Science (New York, N.Y.)*, *212*(4497), pp. 947–949, https://doi.org/10.1126/science.7233191.
- Rodríguez-Valiente, A., Fidalgo, A.R., Villarreal, I.M., García Berrocal, J.R. (2016). Extended High-frequency Audiometry (9000–20000Hz). Usefulness in Audiological Diagnosis, *Acta Otorhinolaryngological (English Edition)*, 67 (1), pp.40-44, DOI:10.1016/j.otoeng.2015.02.001.
- Rodríguez-Valiente, A., Trinidad, A., Garcia Berrocal, J.R., Gorriz, C., & Ramirez Camacho, R., (2014). Extended high-frequency (9–20 kHz) audiometry reference thresholds in 645 healthy subjects, *International Journal of Audiology*, 53 (8), pp.531–545, DOI :10.3109/14992027.2014.893375.
- Samson, Y., Belin, P., Thivard, L., Boddaert, N., Corzier, S., & Zilbovicius, M. (2001). Perception auditive et langage : imagerie fonctionnelle du cortex auditif sensible au langage, *Revue de neurologie*, *157*(8-9), pp. 837-846.
- Schechter, M.A., Fausti, S.A., Rappaport, B.Z., & Frey, R.H., (1986). Age categorization of high-frequency auditory threshold data, *Journal of the Acoustical Society of America,* 79, pp.767–771, DOI:10.1121/1.393466.
- Schroder, M., Burkhardt, F., & Krstulovic, S., (2010). Synthesis of emotional speech. Dans Scherer, K.R., Banziger, T., & Roesch, E. (Eds.), *Blueprint for Affective Computing*. Oxford University Press, pp. 222–231.
- Shadle, C.H., & Damper, R.I. (2001). Prospects for Articulatory Synthesis: A Position Paper, *Image, Speech and Intelligent Systems Research Group, Department of Electronics and Computer Science (ISCA)*, 4th ISCA Workshop on Speech Synthesis, Scotland.
- Shoji, K., Regenbogen, E., Yu, J.D., & Blaugrund, S.M. (1992). High-frequency power ratio of breathy voice, *Laryngoscope*, 102, pp.267–271, DOI:10.1288/00005537-199203000-00007.
- Shroeter, J. (2005). Voice Modification for Applications in Speech Synthesis. *AT&T Labs – Resarch*, pp.1-20.
- Sicard, E., & Menin, A. (1995). Analyse spectrale des sons musicaux et de la parole. *Le Bulletin de l'Union des Physiciens*, (778).
- Singh, S., & Murry, T. (1978). Multidimensional classification of normal voice qualities. *The Journal of the Acoustical Society of America*, 64, pp.81–87, https://doi.org/10.1121/ 1.381958.

- Smith, D.R., & Patterson, R.D. (2005). The interaction of glottal-pulse rate and vocal-tract length in judgements of speaker size, sex, and age. *The Journal of the Acoustical Society of America*, 118(5), pp.3177-3186. https://doi.org/10.1121/1.2047107.
- Stelmachowicz, P.G., Beauchaine, K.A., Kalberer, A., Kelly, W.J., & Jesteadt, W. (1989). High-frequency audiometry: Test reliability and procedural considerations, *Journal of the Acoustical Society of America,* 85, pp.879–887, DOI :10.1121/1.397559.
- Stelmachowicz, P.G., Lewis, D.E., Choi, S. & Hoover, B. (2007), Effect of stimulus bandwidth on auditory skills in normal-hearing and hearing-impaired children, *Ear Hear*, 28 (4), pp.483-494, 10.1097/AUD.0b013e31806dc265.
- Stelmachowicz, P.G., Pittman, A.L., Hoover, B.M., & Lewis, D.E. (2001). Effect of stimulus bandwidth on the perception of /s/ in normal- and hearing-impaired children and adults, *Journal of the Acoustical Society of America*, 110, pp.2183–2190, DOI:10.1121/1.1400757.
- Stevens, C., Lees, N., Vonwiller, J., & Burnham, D. (2005). On-line experimental methods to evaluate text-to-speech (TTS) synthesis: effects of voice gender and signal quality on intelligibility, naturalness and preference, *Computer Speech & Language*, 19 (2), pp. 129-146, https://doi.org/10.1016/j.csl.2004.03.003.
- Sundberg, J. (1987). The science of the singing voice, *Northern Illinois University Press*, 216 pages.
- Sundberg, J. (1995). Vocal fold vibration patterns and phonatory modes, *Folia Phoniatrica Logopedica*, 47 (4), pp.218–228, DOI: 10.1159/000266353.
- Tabet, Y., & Boughazi, M. (2011). Speech synthesis techniques. A survey. *Proceedings of the 7th IEEE International Workshop on System, Signal Processing and their Applications (WOSSPA)*, pp.67-70. https://doi.org/10.1109/WOSSPA.2011 .5931414.
- Terbeek, D. (1977). A Cross-language Multidimensional Scaling Study of Vowel Perception, *UCLA Working Papersin Phoneticss (WPP)*, 37, 267 pages, https://escholarship.org /uc/item/3nx4g138.
- Tervaniemi, M., Just, V., Koelsch, S., Widmann, A., & Schröger, E. (2005). Pitch discrimination accuracy in musicians vs nonmusicians: an event-related potential and behavioral study. *Experimental brain research*, *161*(1), pp.1–10. https://doi.org/10.100 7/s00221-004-2044-5.
- Teston, B. (2004). L'évaluation instrumentale des dysphonies : État actuel et perspectives d'évolution. *In :* Giovanni A, editor. *Le bilan d'une dysphonie : État actuel et perspectives*, Marseille, Solal Collection, pp.105–169.
- Titze, I. (2000). *Principles of voice production*. National Center for Voice and Speech.
- Tullis, T., & Albert, B. (2013). Self-reported metrics. In M. Dunkerley & H. Scherer (Eds.), *Measuring the user experience: Collecting, analyzing, and presenting usability metrics* (2nd ed., pp. 121-161).
- Vaissière, J. (2011). Les organes de la parole, *La phonétique*, pp. 45-56.
- Valencia, N.N., Mendoza, L.E., Mateo, R.I., & Carballo, G.G. (1994). High-frequency components of normal and dysphonic voices, *Journal of Voice*, 8, pp.157–162, https://doi.org/10.1016/S0892-1997(05)80307-8.

- Van Dommelen, W.A. (1990). Acoustic parameters in human speaker recognition, *Language and Speech*, 33, pp.259–272, DOI:10.1177/002383099003300302.
- Van Niekerk, D. R., Xu, A., Gerazov, B., Krug, P. K., Birkholz, P., & Xu, Y. (2022). Exploration strategies for articulatory synthesis of complex syllable onsets. *arXiv preprint arXiv:2204.09381*.
- Vitela, A.D., Monson, B.B., & Lotto, A.J. (2015). Phoneme categorization relying solely on high-frequency energy. *The Journal of the Acoustical Society of America*, 137(1), pp.65-70. https://doi.org/10.1121/1.4903917.
- Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Eje Henter, G., Le Maguer, S., Malisz, Z., Székely, É., Tånnander, C., & Voße, J. (2019). Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program., *Proc. 10th ISCA Speech Synthesis Workshop*, pp.105-110, DOI: 10.21437/SSW.2019-19.
- Walden, B.E., Montgomery, A.A., Gibeily, G.T., Prosek, R.A., & Schwartz, D.M. (1978). Correlates of psychological dimensions in talker similarity, *Journal of Speech and Hearing Research*, 21, pp.265–275, DOI: 10.1044/jshr.2102.265.
- Wright R. (2019), Voice quality types and uses in north American English, https://journals.openedition.org/anglophonia/1952.
- Yamasaki, R., Montagnoli, A., Murano, E. Z., Gebrim, E., Hachiya, A., da Silva, J. V. L., ... & Tsuji, D. (2017). Perturbation measurements on the degree of naturalness of synthesized vowels. *Journal of Voice*, *31*(3), 389-e1.
- Young, A.W., & Bruce, V. (2011). Understanding person perception. *British Journal of Psychology*, *102*, pp.959–974. doi:10.1111/j.2044-8295.2011.02045.x.
- Zadeh, L. M., Silbert, N. H., Sternasty, K., Swanepoel, D.W., Hunter, L. L., & Moore, D.R. (2019). Extended high-frequency hearing enhances speech perception in noise. *Proceedings of the National Academy of Sciences*, 116(47), pp.1-7. https://doi.org/10.1073/pnas.1903315116.

IX. ANNEXES

Annexe 1 : Questionnaire anamnestique

QUESTIONNAIRE ANAMNESTIQUE

Initiales (prénom, nom) :

Genre : ☐ Homme ☐ Femme ☐ Autre

Age au moment du test :

- En quelle année d'études êtes-vous inscrit actuellement ?

  _____
  _____

- Dans quelle filière êtes-vous inscrit à l'Université de Liège ?

  _____
  _____

- Quelle est votre langue maternelle ?

  _____
  _____

- Êtes-vous bilingue ? ☐ Oui ☐ Non

  Si OUI, quelles sont les langues que vous connaissez (qu'elles soient parlées ou uniquement comprises) ?

  _____
  _____
  _____
  _____

- Avez-vous déjà rencontré des problèmes d'audition (par exemple, des otites à répétition) ?

  ☐ Oui ☐ Non

  Si OUI, quel(s) problème(s) et à quel âge ?

  _____
  _____
  _____
  _____

Si OUI, quelle(s) intervention(s) médicale(s) avez-vous subie(s) afin de corriger vos problèmes d'audition (par exemple, une pose de drains trans-tympaniques) ?

_____
_____
_____
_____

- Rencontrez-vous, en ce moment, des problèmes d'audition/d'oreille ? ☐Oui ☐Non

Si OUI, lesquels ?

_____
_____
_____
_____

- Portez-vous un appareil auditif ? ☐ Oui ☐Non

AUTRES DONNÉES :

- Possédez-vous des connaissances dans le domaine de la synthèse de la parole ? ☐Oui ☐Non
  Si OUI, pouvez-vous indiquer où et quand vous avez acquis ces connaissances ?

_____
_____
_____
_____

En quoi consistent vos connaissances ?

☐Connaissances théoriques ☐Connaissances pratiques ☐Autre

Si vous avez coché « Autre », précisez :

_____
_____
_____
_____

Si NON :

Avez-vous déjà entendu parler de la synthèse de la parole ? ☐Oui ☐Non

Annexe 2 : Rapport d'étalonnage de l'audiomètre Madsen Itera II

Sonova
HEAR THE WORLD

Client Nr.

Bon Nr. Audiologie

Client Name CHU Liège.

Address

Date 22/03/2021

Repair ☐
Remote Intervention ☐
Installation ☐
Sales Support ☐
Calibration ☒
Quotation ☐
Delivery Note ☐

Order

Customer Nr.

Brand Madsen

Type ITERA

Class

SN 202163
2004

Description Révision générale.
Etalonnage. + 12500 Hz.

Frequency Hz	125	250	500	750	1000	1500	2000	3000	4000	6000	8000	H.F.	M.F.
	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	12500Hz	

Calibration Tone	Headph. L.F.		Bone	Masking		Insert Ph.		Masking Ins.		White Noise		High Freq.	
	L	R		L	R	L	R	L	R	L	R	L	R
	✓	✓		✓	✓					✓	✓	✓	✓

Calibration Speech	Headphones		Bone	Free Field		Calibration Speech Noise	Headphone		Free Field		Headph. H.F.	
	L	R		1	2		L	R	1	2		

Qty	Spare Parts / Accessories	Unit Price	Total Price

Total Spare Parts / Accessories	
Total Work	
Calibration ☒ Bone ☐ Speech ☐ Free Field ☐ High Frq. ☒ Ins. Ph. ☐ Multi Frq. ☐ HIT ☐ REM ☐	
Transport	
Total (excl.TVA)	

Remarks

For approval	Engineer	Standards
Name	G. DELRUE	ISO 389-1 / ISO 389-2 / ISO 389-3 ISO 389-4 / ISO 389-5 / ISO 389-6 ISO 389-7 ISO 389-8 / ISO 389-9
Signature Del.	On Site ☒ Labo ☐	Warranty ☐ Contract ☐

Sonova Retail Belgium | Stationsstraat 22 | B-1702 Groot-Bijgaarden | tel 02 700 77 05 | audiologie@sonova.be

Annexe 3 : Audiogramme utilisé pour chaque participant

AUDIOGRAMME

Initiales (prénom, nom) :

Genre : ☐ Homme ☐ Femme ☐ Autre

Age au moment du test :

Date :

OREILLE DROITE

OREILLE GAUCHE

Annexe 4 : Seuils auditifs obtenus pour chaque participant

Participant	Oreille droite						Oreille gauche					
	0.5 kHz	1 kHz	2 kHz	4 kHz	8 kHz	12.5 kHz	0.5 kHz	1 kHz	2 kHz	4 kHz	8 kHz	12.5 kHz
1	10	10	5	5	5	0	10	10	5	5	5	5
2	5	5	5	5	5	5	5	5	5	5	5	5
3	5	10	5	5	5	5	5	10	10	5	10	5
4	10	15	10	5	0	0	10	15	10	5	5	5
5	5	15	10	5	5	0	5	10	10	5	5	5
6	5	10	10	5	5	5	5	10	10	10	5	5
7	5	10	10	5	0	5	10	5	10	0	5	5
8	5	5	5	5	10	10	10	5	5	0	5	10
9	5	10	10	0	5	5	10	10	5	0	5	0
10	10	10	5	0	5	0	5	0	0	5	5	0
11	10	10	5	0	5	5	10	10	5	5	5	10
12	10	5	10	0	5	5	10	10	0	0	5	5
13	5	10	5	5	5	5	5	10	10	5	5	5
14	10	0	5	0	0	0	10	10	0	0	0	5
15	10	10	10	5	0	0	10	5	0	5	0	5
16	10	10	10	0	5	5	10	5	0	5	0	5
17	10	5	10	0	5	0	10	10	0	5	5	0
18	10	10	10	5	5	5	10	10	5	0	0	5
19	10	5	5	0	5	0	5	10	5	5	0	0
20	10	10	5	0	5	10	5	10	5	0	5	5
21	10	5	0	0	0	0	5	5	0	0	0	0
22	10	5	10	0	0	5	5	5	0	0	0	5
23	10	10	5	5	5	5	10	10	0	0	5	5
24	5	0	5	10	10	15	5	5	5	10	10	10
25	5	5	0	0	0	0	5	5	5	0	0	5
26	10	15	10	5	0	5	10	10	5	5	0	5
27	10	10	5	5	5	10	15	10	5	5	5	5
28	5	0	0	0	0	5	10	10	5	0	0	0
29	10	10	5	0	0	0	15	10	0	0	0	0
30	10	10	5	5	10	0	5	10	0	5	5	0
31	10	15	5	5	0	0	15	15	5	0	0	0
32	10	10	5	5	0	5	15	15	15	10	0	0
33	10	10	5	5	5	0	5	10	5	10	5	5
34	15	10	5	10	5	5	10	10	5	15	5	5
35	10	5	5	0	0	5	10	10	5	0	5	5
36	10	15	0	5	5	10	10	15	10	10	10	10
37	0	0	0	0	5	10	5	5	0	0	5	0
38	5	10	5	15	10	10	5	10	5	5	0	5
39	10	15	5	10	0	0	5	15	10	5	10	5
40	10	10	0	5	0	0	5	5	5	10	0	5

Annexe 5 : Protocole du testing

Protocole expérimental – Fontaine Camille

1. Étapes à suivre avec le participant

- Aller chercher le participant à l'endroit du rendez-vous (5 minutes)

  - L'inviter à respecter les gestes barrières et à se désinfecter les mains à l'aide du gel hydroalcoolique présent sur place.

  - L'accompagner jusqu'à la cabine audiométrique.

- Complétion des documents administratifs (10 minutes)

  - Lui fournir les papiers à compléter (formulaire de consentement, formulaire d'informations aux volontaires, questionnaire anamnestique).

  - Lui demander de compléter et de signer les documents.

  - Lui demander de mettre son téléphone et son ordinateur en mode avion.

- Réaliser l'audiométrie tonale (25 minutes).

  - Installer correctement le participant :
    - Position confortable ?
    - Retirer les lunettes et/ou les boucles d'oreilles si besoin.
    - Interdiction de manger et/ou de boire durant le testing.
    - Regarder droit devant soi et se concentrer.

  Donner la consigne :

  - « *Vous allez écouter des sons dans le casque. Lorsque vous entendez un son, levez la main. Si vous n'entendez pas de son, ne faites rien. Nous allons réaliser tout le testing sur l'oreille gauche, et puis nous ferons la droite. Je vais vous aider à positionner le casque.* ».

  - Positionner le casque :
    - Veiller à l'absence de cheveux entre les oreilles et le casque.
    - Placer le côté rouge sur l'oreille droite et le côté bleu sur l'oreille gauche.
    - Réaliser un essai en donnant la consigne suivante : « *Avant de commencer, nous allons faire un essai. Est-ce que vous entendez un son ? Dans quelle oreille entendez-vous ce son ?* ».

- Placer une cloison (la valise de l'audiomètre) entre le participant et l'audiomètre afin que le participant ne remarque pas quand le son est lancé. Veiller à ne bouger que les doigts sur l'audiomètre et non les bras. Garder un visage neutre.

- Commencer à tester à 25 dB à 1000 Hz.
    - Descendre par pas de 5 dB jusqu'à ce que le participant n'entende plus.
    - Lorsque le stade où le participant n'entend plus est atteint, remonter de 5 dB pour voir si le participant entend à nouveau le signal.
    - Si la participant entend à nouveau, redescendre de 5 dB pour vérifier qu'il n'entend plus.
    - Recommencer cette procédure pour toutes les autres fréquences.

    Ordre des fréquences testées : 500 Hz, 1000 Hz, 2000 Hz, 4000 Hz, 8000 Hz, 12500 Hz.
    → Avec l'audimètre MADSEN Itera II et le casque Sennheiser HDA 300.

- Retirer l'audiomètre de la table, le ranger sur le côté.

o Réaliser la tâche 1 : Comparaison par paires (25 minutes).

- Brancher le haut-parleur et/ou le casque à l'ordinateur.

- Mettre le son de l'ordinateur au maximum et le volume du diffuseur sur 1.5.

Donner la consigne suivante : « *Vous allez écouter 2 sons, et vous devez indiquer quel est le son qui vous parait le plus naturel. Pour écouter les sons, vous devez cliquer sur le bouton « son 1 » ou sur le bouton « son 2 ». Ensuite vous choisirez le son qui vous parait le plus naturel en cliquant sur le bouton « le son 1 est plus naturel » ou « le son 2 est plus naturel ». Pour passer à la paire suivante, vous devez cliquer sur le bouton « paire suivante ». Vous pouvez écouter les sons autant de fois que vous le souhaitez. Néanmoins, lorsque vous aurez cliqué sur le bouton « paire suivante », vous ne pourrez plus revenir en arrière. Nous allons d'abord faire un entrainement avec 2 paires de sons. Durant le testing, vous aurez 120 paires de sons à comparer. Après l'essai, n'hésitez pas à poser vos éventuelles questions ».*

- Réaliser un essai.

- Lancer la tâche 1.

- Encoder le numéro du participant en respectant l'anonymisation du participant.

- Enregistrer le fichier.

- Réaliser une pause à l'extérieur de la cabine audiométrique si le participant le souhaite (5 minutes).

- Réaliser la tâche 2 : Évaluation du degré de naturel (20 minutes).

  - Vérifier que le son de l'ordinateur est toujours au maximum et que le haut- parleur est toujours bien branché et réglé sur 1.5.

  - Donner la consigne suivante : « *Vous allez désormais écouter 120 sons, l'un à la suite de l'autre, et vous allez devoir évaluer son degré de naturel sur base d'une échelle allant de 0 « pas du tout naturel » à 100 « tout à fait naturel ». Pas du tout naturel signifie que le son que vous écoutez ressemble à une voix artificielle, éloignée d'une voix réelle. Tout à fait naturel signifie que le son que vous écoutez ressemble à une voix réelle. L'écoute du son se réalise de la même façon que dans la tâche 1. Quant au degré de naturel, vous le déterminerez en déplaçant le curseur se situant en bas de l'interface. Lorsque vous avez cliqué sur le bouton « paire suivante », le son suivant se lance automatiquement, mais vous pouvez tout de même écouter ce son autant de fois que vous le souhaitez en cliquant sur le bouton « écouter ». Néanmoins, une fois que vous avez cliqué sur le bouton « suivant », vous ne pouvez pas revenir en arrière. Nous allons réaliser un essai, après lequel vous pourrez poser vos éventuelles questions.* ».

  - Lancer la tâche 2.

  - Encoder le numéro du participant en respectant l'anonymisation du participant.

  - Enregistrer le fichier.

- Laisser le participant partir et le remercier pour sa participation à l'expérience.

- Donner les 10 euros de dédommagement et faire signer l'attestation de paiement.

- Procéder à une désinfection du local et à une aération de celui-ci.

2. Documents à apporter par passation

- Formulaire de consentement (2x)
- Formulaire d'informations aux volontaires (1x)
- Questionnaire anamnestique (1x)
- Audiogramme (1x)
- Attestation de paiement (1x)

Annexe 6 : Interaction entre le type de phonème et le genre de voix de synthèse

D'un point de vue descriptif, la *Figure 22* illustre les scores obtenus pour l'aspect naturel de chaque phonème, selon le genre de voix de synthèse (féminin ou masculin).

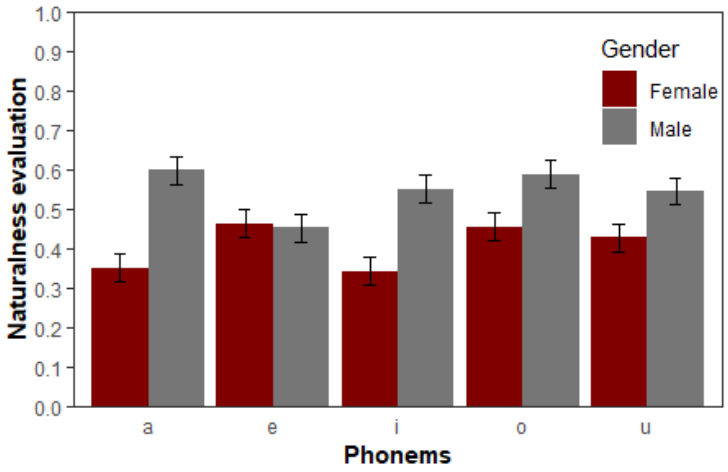

Figure 22 : Score moyen attribué pour l'aspect naturel de chaque phonème en fonction du genre de voix de synthèse

Les phonèmes [a], [i], [o], [u] générés avec la voix masculine sont pratiquement toujours jugés comme plus naturels que la voix féminine, à l'exception du [e]. La taille de la différence entre la voix masculine et la voix féminine varie selon le phonème. Pour le phonème [a], la différence est la plus importante, tandis que pour le phonème [e], la différence est la plus petite. D'un point de vue statistique, plusieurs différences significatives ont été mises en évidence au travers de contrastes linéaires.

Pour le phonème [a], il existe une différence significative entre la voix de synthèse féminine et la voix de synthèse masculine ($Z = -17.290$, $p < .001$), il en va de même pour le phonème [i] ($Z = -14.635$, $p < .001$), pour le phonème [o] ($Z = -9.474$, $p < .001$) et pour le phonème [u] ($Z = -6.425$, $p < .001$). Concernant le phonème [e], aucune différence significative n'a été mise en évidence ($Z = 0.825$, $p > .05$). Les résultats des contrastes linéaires confirment qu'il existe une **différence concernant l'aspect naturel selon le phonème ([a], [e], [i], [o] ou [u]) et selon le genre de voix de synthèse (féminin ou masculin)**. Cette analyse est une analyse complémentaire pour laquelle nous n'avions pas émis d'hypothèse ou de question de recherche. Elle ne sera pas traitée dans la discussion.

Annexe 7 : Résultat de l'interaction degré de réalisme du modèle acoustique * type de phonème (2.1.5.1)

Phonème – Modèle	Phonème – Modèle contrasté	Score Z	Probabilité (>|Z|)
1D - [a]	BWE [a]	8.561	**0.000*** **
1D - [e]	BWE [e]	8.397	**0.000*** **
1D - [i]	BWE - [i]	4.851	**0.000*** **
1D - [o]	BWE - [o]	- 0.725	1
1D - [u]	BWE - [u]	- 4.941	**0.000*** **
1D - [a]	3D - [a]	- 1.218	1
1D - [e]	3D - [e]	- 0.009	1
1D - [i]	3D - [i]	0.277	1
1D - [o]	3D - [o]	- 4.867	**0.000*** **
1D - [u]	3D - [u]	- 6.666	**0.000*** **
BWE - [a]	3D - [a]	- 9.773	**0.000*** **
BWE - [e]	3D - [e]	- 8.429	**0.000*** **
BWE - [i]	3D - [i]	- 4.584	**0.000*** **
BWE - [o]	3D - [o]	- 4.142	**0.001** **
BWE - [u]	3D -[u]	- 1.699	1

Résumé

Introduction et objectifs : La parole de synthèse peut-être réalisée via diverses méthodes, notamment la synthèse articulatoire. Il existe différentes modélisations physiques : le modèle acoustique unidimensionnel (1D), le modèle acoustique tridimensionnel (3D) et le modèle d'algorithme d'extension (BWE). Le modèle 3D semble offrir la parole la plus naturelle (Gully, 2017). D'une part, il se base sur la forme précise du tractus vocal, générant des simulations acoustiques proches de la réalité, et d'autre part, il permet de modéliser correctement les hautes fréquences (HF) (> 5kHz) (Arnela et al., 2019 ; Freixes et al., 2018). Longtemps mises de côté dans les recherches sur la perception de la parole, ces HF connaissent un nouvel intérêt depuis plusieurs années, car elles semblent jouer un rôle important dans l'aspect naturel de la parole (Vitela et al., 2015 ; Monson & Caravello, 2019 ; Boyd-Pratt & Donnai, 2020 ; Birkholz & Drechsel, 2021). Ce mémoire s'inscrit dans un projet de développement d'un outil de synthèse articulatoire à large bande, dont l'aspect se veut le plus naturel possible. Notre objectif est de déterminer, pour la synthèse articulatoire, comment les différents modèles : 1D, 3D et BWE, impactent la perception de l'aspect naturel de la parole chez les jeunes adultes. Méthodologie : Après avoir rempli un questionnaire anamnestique et passé une audiométrie tonale, 40 participants ont réalisé deux tâches expérimentales. La première tâche était une comparaison par paires, qui avait pour but de comparer l'aspect naturel des différents stimuli deux à deux. La seconde tâche était une évaluation de l'aspect naturel des stimuli, à l'aide d'une échelle métrique allant de 0 (pas du tout naturel) à 100 (totalement naturel). Ces tâches nous ont permis de répondre à trois hypothèses concernant le degré de réalisme physique des modèles acoustiques, et d'investiguer différentes questions de recherche concernant le genre de la voix de synthèse, la qualité vocale, et le type de phonème, et la fiabilité inter-juges. Résultats et conclusions : Les deux tâches expérimentales ont permis de mettre en avant plusieurs effets significatifs. Un effet significatif du modèle acoustique a été trouvé, de façon générale, le modèle 3D est plus naturel. Un effet significatif du type de phonème a montré que le degré de naturalité dépend du phonème. Une interaction a été trouvée entre le modèle acoustique et le type de phonème, révélant que l'aspect naturel des modèles diffère selon le type de phonème. Seule la seconde tâche expérimentale a permis de mettre en lumière un effet significatif du genre de la voix de synthèse, indiquant que la voix de synthèse masculine parait plus naturelle que la féminine. Ce mémoire a cherché à explorer le rôle des HF dans la perception des phonèmes selon différents degrés de réalismes physiques de modèles acoustiques.