Dans quelle mesure la complexité physique en hautes fréquences de modèles de synthèse de la parole différents impacte-t-elle la perception de phonèmes ?

Auteur : Hoornaert, Céleste
Promoteur(s) : Remacle, Angélique; 13307
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 : 2020-2021
URI/URL : http://hdl.handle.net/2268.2/13467

Université de Liège

Faculté de Psychologie, Logopédie et Sciences de l'Education

Dans quelle mesure la complexité physique en hautes fréquences de modèles de synthèse de la parole différents impacte-t-elle la perception de phonèmes ?

Mémoire présenté en vue de l'obtention du grade de

Master en Logopédie, à finalité spécialisée en voix

Promoteur : Angélique REMACLE

Co-promoteur : Rémi BLANDIN

Céleste HOORNAERT

Année académique 2020-2021

REMERCIEMENTS

A mes promoteurs, Madame Angélique Remacle et Monsieur Rémi Blandin,
Pour votre disponibilité, votre écoute, votre compréhension,
Pour votre humanité dans les moments plus difficiles,
Pour vos nombreux conseils tout au long de ces deux années de travail.

Aux lecteurs de mon mémoire, Vincent Didone et Morgane Warnier,
Pour l'aide apportée pour la réalisation des analyses statistiques,
Pour l'intérêt que vous avez porté à ce mémoire.

A Monsieur Xavier Kaiser,
Pour votre disponibilité, votre aide et vos réponses à mes questions,
Pour le développement de l'audiomètre qui a permis de mener à bien l'étude.

A tous les participants des pré-tests et des testings,
Pour vous être intéressés à mon mémoire et son objet d'étude,
Pour m'avoir parfois aidée à trouver d'autres participants,
Pour m'avoir montrée la solidarité estudiantine.

A ma famille, en particulier ma mère, ma grand-mère, Raynald, Anthony et Ori,
Pour m'avoir supportée dans mes grands moments de stress,
Pour m'avoir écoutée lors de mes baisses de moral, au téléphone ou en face,
Pour m'avoir encouragée tout au long de mon parcours académique, bien avant mon entrée à l'université,
Pour avoir cru en moi,
Vous avez été ma plus grande motivation et mon plus grand soutien. Merci.

A vous tous qui m'avez aidée de près ou de loin à la conception de ce travail, je tiens à vous remercier le plus chaleureusement possible.

TABLE DES MATIERES

LISTE DES ABREVIATIONS

f_0	Fréquence fondamentale
HF	Hautes fréquences
HFE	Hautes fréquences étendues
CVC	Consonne-voyelle-consonne
VCV	Voyelle-consonne-voyelle
CV	Consonne-voyelle
VC	Voyelle-consonne
OEA	Otoémissions acoustiques
kHz	Kilohertz
dB	Décibel
1D	Unidimensionnel
3D	Tridimensionnel
ICC	Coefficient de corrélation intra-classe

LISTE DES TABLEAUX ET FIGURES

TABLEAUX

FIGURES

I. INTRODUCTION GENERALE

Associées à une importance mineure pour la perception de la parole, les hautes fréquences (> 5 kHz) ont été ignorées dans la majorité des recherches (Monson & Caravello, 2019 ; Vitela et al., 2015). Une remise en question de leur faible contribution pour percevoir la parole a abouti à un regain d'intérêt dans les années 2010 (Boyd-Pratt & Donai, 2020). Jusqu'alors, aucune étude n'était parvenue à démontrer leur inutilité perceptive (Monson, Hunter et al., 2014). Certains auteurs ont d'ailleurs mis en évidence la capacité humaine à percevoir les hautes fréquences dans la parole et ont supposé qu'elles contiennent de l'information pertinente dans le signal de parole (Monson et al., 2011 ; Monson & Caravello, 2019). Birkholz et Drechsel (2021) sont allés plus loin en suggérant le potentiel rôle des hautes fréquences pour produire un signal de parole plus naturel. Nous avons souhaité investiguer cette question plus en profondeur à l'aide de la synthèse de parole.

Diverses méthodes de synthèse de parole existent, dont la synthèse articulatoire. Elle consiste en une représentation modélisée de l'appareil phonatoire humain (Birkholz et al., 2006). Son principal intérêt est d'étudier le lien entre le fonctionnement de la production de parole et la perception de parole, en investiguant l'influence de changements anatomiques du tractus vocal sur les signaux acoustiques générés (Huang, 2001). La synthèse articulatoire n'est pour le moment pas adaptée pour un usage commercial, comme dans les applications utilisant de la synthèse vocale (GPS, Google Home, etc.) (Lukose & Upadhya, 2017). La littérature scientifique met en évidence les avantages à l'employer dans les études perceptives (Birkholz et al., 2017) puisqu'elle n'est pas influencée par des effets liés au locuteur, eux-mêmes susceptibles d'influencer la production du signal de parole (Delvaux & Pillot-Loiseau, 2019). Fondée sur deux types de modélisation physique, plusieurs auteurs soulignent le plus grand réalisme des simulations acoustiques générées par la modélisation tridimensionnelle (3D), grâce à la prise en compte précise de la forme du tractus vocal (Arnela et al., 2019 ; Freixes et al., 2018). Bien que son usage soit encore peu répandu, l'emploi d'un modèle 3D avec la synthèse articulatoire semble prometteur pour explorer la potentielle contribution des hautes fréquences dans l'aspect naturel de la parole (Arnela et al., 2019 ; Freixes et al., 2018 ; Monson, Hunter et al., 2014). A ce jour, aucune étude ne semble avoir évalué les différences perceptives générées par l'usage des modèles 1D et 3D auprès d'auditeurs réels.

Une meilleure connaissance du rôle des hautes fréquences pour la perception de la parole sous-tend plusieurs enjeux. Dans le cas où l'intégration des hautes fréquences dans la synthèse de la parole s'avère inutile, nous pourrions nous satisfaire de modélisations acoustiques simples pour synthétiser la parole. Dans le cas contraire, l'inclusion des hautes fréquences impliquerait le développement de modèles acoustiques plus complexes pour générer de la parole synthétique. Un tel résultat soulèverait des enjeux, notamment pour les logopèdes et les audiologistes qui travaillent auprès de patients déficients auditifs (Boyd-Pratt & Donai, 2020). Il permettrait de les sensibiliser aux multiples implications des hautes fréquences dans la perception de la parole, aux éventuels enjeux de dépister précocement une perte auditive dans les hautes fréquences et d'assurer une suivi thérapeutique adapté pour le patient (Boyd-Pratt & Donai, 2020 ; Hunter et al., 2020).

Ce mémoire s'inscrit dans un projet de développement d'une synthèse articulatoire à large bande. Il est le fruit d'une collaboration entre Angélique Remacle, chercheuse dans l'Unité Logopédique de la Voix de l'ULiège, et Rémi Blandin, chercheur dans l'Institut d'acoustique et de la parole de l'Université technique de Dresde. Son objectif est de mieux comprendre et définir le lien entre la perception de la parole et les aspects physiques et acoustiques liés à sa production dans l'entièreté du registre fréquentiel audible (0.02 à 20 kHz). Pour répondre à cette problématique, deux types de modélisation physique des hautes fréquences (> 5 kHz) sont utilisés : la modélisation tridimensionnelle (3D) et la modélisation unidimensionnelle (1D). Au regard des données relevées dans la littérature scientifique, une sensibilité auditive entre ces deux degrés de réalisme dans la synthèse des HF devrait être objectivée. En outre, les phonèmes générés avec le modèle 3D devraient être considérés comme plus naturels que ceux générés avec le modèle 1D, compte-tenu son plus grand réalisme acoustique.

Plusieurs sections composent ce mémoire. La première correspond à l'introduction théorique qui aborde trois thématiques principales relatives au projet : l'analyse perceptive de la parole, les hautes fréquences et les méthodes de synthèse de la parole. La seconde section traite les objectifs et hypothèses formulés à partir des données relevées dans la littérature. La troisième section expose la méthodologie mise en place. La quatrième section présente l'ensemble des résultats obtenus grâce aux analyses statistiques. Ces résultats sont discutés et interprétés dans la dernière section, qui aborde également les limites, biais méthodologiques et perspectives envisagées pour améliorer ce travail.

II. REVUE DE LA LITTERATURE

1. Analyse perceptive de la parole

La parole est un instrument que nous utilisons si naturellement que nous en oublions parfois toute sa complexité. Résultant de la manipulation de la pression d'air générée par le système respiratoire, elle consiste en un enchaînement rapide de mouvements par les organes de la parole[1] et implique la coarticulation de différents phonèmes (Baken & Orlikoff, 2000 ; Castellengo, 2015). Bien qu'étroitement liées, les notions de voix et de parole se distinguent. La parole peut ou non s'accompagner de voix, c'est-à-dire qu'elle se définit par des sons voisés et des sons non voisés, tels que des écoulements d'air. Générée par l'appareil phonatoire, la voix consiste en la mise en vibration des plis vocaux qui convertit le flux d'air provenant des poumons en énergie acoustique (Kreiman & Sidtis, 2011). Ce mémoire se centre sur la perception de la parole, qui correspond à la capacité d'extraction des unités linguistiques élémentaires (phonèmes) et de leurs caractéristiques acoustiques distinctives dans le signal de parole (Diehl et al., 2004).

1.1. Corrélats cérébraux

Les voix sont une composante spéciale pour le cerveau humain (Belin et al., 2011). D'après Belin et al. (2004), la voix est probablement le son le plus représenté au sein de notre environnement auditif.

L'apparition de techniques d'imagerie cérébrale a permis d'objectiver le lien étroit entre la voix et l'humain à travers la découverte des corrélats cérébraux. Ces derniers correspondent aux régions cérébrales exclusivement dédiées au processus de perception de la voix. L'écoute de stimuli vocaux active des zones cérébrales particulières, localisées bilatéralement et longeant les parties moyennes et antérieures du gyrus temporal supérieur. Ces zones sont nommées les **aires temporales de la voix** (Belin et al., 2011) (voir figure 1).

[1] Les organes de la parole sont les poumons, le larynx, la langue, les lèvres et le voile du palais (Vaissière, 2011)

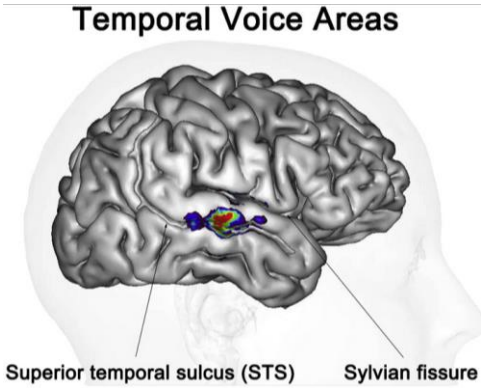

Figure 1 : Aires temporales vocales dans le cerveau adulte situées sous la scissure de Sylvius, le long du sillon temporal supérieur (Belin & Grosbras, 2010)

1.1.1. Implication bilatérale du lobe temporal supérieur postérieur

De nombreuses études ont tenté de définir précisément la localisation de ces aires temporales vocales, sans qu'il n'y ait de réel consensus au départ. Dans ce contexte controversé, Boatman et al. (1995) ont cherché à déterminer l'organisation cérébrale sous-jacente au traitement de la parole. L'interférence électrique directe a été utilisée chez trois patients ayant subi une intervention chirurgicale dans le but de guérir leur épilepsie. Elle se définit par l'envoi de décharges électriques qui perturbent l'activité électrique neuronale normale. Elle avait lieu durant la passation de 3 tâches : une identification de voyelles et de consonnes (qui consiste à choisir le stimulus correct parmi des distracteurs), une discrimination de voyelles (qui consiste à distinguer deux stimuli l'un de l'autre) et une épreuve de compréhension de phrases (qui consiste à répondre à des instructions de plus en plus complexes à travers une manipulation de jetons). L'objectif était d'observer les effets d'une stimulation électrique sur l'anatomie fonctionnelle de la perception de parole. Les résultats ont révélé l'existence de régions corticales cruciales pour son traitement. La stimulation électrique d'une zone cérébrale particulière engendrait l'échec des patients dans les trois épreuves. Il s'agit du **lobe temporal supérieur postérieur**. En 2000, Hickok et Poeppel ont constaté que la présence de lésions focales au sein de l'aire de Broca, du gyrus temporal médian ou encore du gyrus supramarginal n'engendrait pas de difficultés de compréhension. Ces structures étant adjacentes au lobe temporal supérieur postérieur, les résultats convergent vers l'idée qu'il s'agit d'une région critique pour la perception de la parole et de la voix. Sa spécificité se marque par son absence d'activation pour les bruits environnementaux, ce qui la distingue du cortex auditif primaire (Samson et al., 2001).

Un débat subsistait toujours quant à l'implication bilatérale ou unilatérale du lobe temporal supérieur postérieur. La stimulation électrique unihémisphérique dans l'étude de Boatman et al. (1995) ne permettait pas de répondre à cette question. La revue de la littérature de Hickok et Poeppel (2000) a mis à jour l'activation bilatérale de cette zone cérébrale lors de la perception de la parole. Selon ces auteurs, le fait d'écouter passivement des stimuli vocaux activerait systématiquement le lobe temporal supérieur postérieur bilatéralement.

Bien que l'implication bilatérale fasse aujourd'hui consensus, il semble que les deux hémisphères cérébraux n'interviennent pas identiquement lors de l'analyse perceptive de la parole. Kreiman et Sidtis (2011) indiquent que l'hémisphère gauche serait destiné à traiter les aspects phonétiques, phonologiques, morphologiques et syntaxiques. Il s'occuperait d'éléments cruciaux pour la parole, relatifs au séquençage et au timing. Le versant pragmatique, en lien avec les émotions, la familiarité et les inférences, serait traité par l'hémisphère droit.

1.2. Modèle de perception de la voix

La découverte du lobe temporal supérieur postérieur bilatéral en tant que régisseur de la perception vocale a mis en lumière le caractère particulier de la voix pour l'humain. Cependant, qu'implique ce processus de perception vocale et quelles informations est-il susceptible de nous fournir ?

Belin et al. (2004) ont modélisé le traitement cortical précis nécessaire à l'analyse d'une production vocale (voir figure 2). Il convient de préciser que nous n'aborderons pas le lien émis avec le phénomène de perception faciale car ce domaine ne s'inscrit pas dans le cadre de notre recherche.

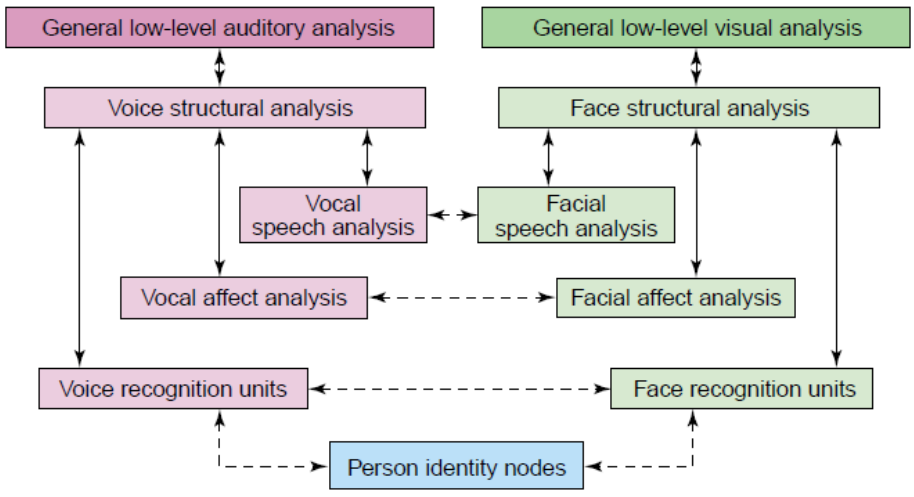

Figure 2 : Modèle de la perception vocale (Belin et al., 2004)

D'après le modèle de Belin et al. (2004), lorsque nous percevons une voix, un traitement de bas niveau au sein des cortex auditifs primaires (*general low-level auditory analysis*) a lieu. Le stimulus vocal se voit ensuite encodé en terme structurel (*voice structural analysis*). Cet encodage se décompose en trois catégories fondées sur les différents aspects véhiculés par l'information. La première renvoie à l'analyse de la parole (*vocal speech analysis*). Elle se fonde sur divers indices que sont les caractéristiques invariantes, comme le timbre du sujet, et les informations dynamiques, telles que l'accent de l'individu. Ces caractéristiques invariantes sont explorées plus en détails dans le point 1.4.. La seconde catégorie est liée à l'analyse de l'aspect affectif (*voice affect analysis*). La dernière fait référence à la reconnaissance des unités vocales (*voice recognition units*), qui aboutit à la construction de l'identité de la personne (*person identity nodes*) (Belin et al., 2004).

Le modèle suggère l'existence de catégories indépendantes les unes des autres, entre lesquelles une relation interactionnelle subsiste. Cette interaction aboutit à l'élaboration de représentations de plus en plus abstraites, qui elles-mêmes conduisent à la construction de l'identité de l'individu entendu. Plus le traitement est de haut niveau, plus il est abstrait et indépendant. A titre illustratif, la reconnaissance de l'identité du sujet ne dépend pas du contenu de sa production vocale, ni de l'état affectif perçu dans sa voix. Cette proposition peut être liée avec l'étude précédemment citée de Boatman et al. (1995). Les résultats de ces auteurs indiquaient déjà une indépendance entre les processus perceptifs de haut et de bas niveau. Les participants rataient parfois uniquement la tâche de compréhension de phrases sans que leur performance ne soit altérée pour les épreuves de discrimination et d'identification. Par ailleurs, l'analyse de certaines erreurs commises lors de la discrimination a montré des effets liés au statut lexical et à la proximité phonologique entre les syllabes. Ce constat laisse supposer que le traitement de haut niveau exerce une influence sur le niveau inférieur, suggérant la présence d'une bidirectionnalité entre les processus. Cet aspect bidirectionnel est concordant avec le modèle de Belin et al. (2004) qui l'ont représenté par leurs flèches dans la figure 2.

1.2.1. Variabilité interindividuelle de l'évaluation perceptive de la voix

Plusieurs indices sont utilisés pour aboutir à la construction de nœuds identitaires sur l'individu, étape finale du modèle de Belin et al. (2004). Toutefois, il semblerait que les stratégies perceptives, employées pour évaluer la voix, diffèrent entre les auditeurs.

La qualité vocale peut être citée en tant qu'indice perceptif. Elle est définie par l'impression perceptive générée par l'écoute du signal acoustique et est dépendante de deux

sources (Kreiman & Sidtis, 2011). D'une part, les différences organiques peuvent être nommées. Elles sont fondées sur la configuration physiologique de l'individu. Relativement invariantes, leur stabilité à l'âge adulte leur confère une certaine fiabilité afin de statuer sur l'identité, le genre et l'âge de l'individu. D'autre part, les différences relatives aux comportements habituels, tels que l'accent, peuvent être mentionnées. Elles sont des indices quant à l'appartenance à une communauté précise et reflètent certaines caractéristiques personnelles importantes.

D'après Delvaux et Pillot-Loiseau (2019), la qualité vocale est multidimensionnelle. Des interactions entre la tâche expérimentale, le locuteur et des facteurs propres à l'auditeur influenceraient la perception de cette qualité vocale. Plomp (1976, cité par Kreiman & Sidtis, 2011) s'est davantage centré sur les dimensions acoustiques de la qualité vocale, telles que l'enveloppe spectrale, la fréquence fondamentale, l'amplitude du signal, le caractère périodique ou apériodique et l'importance de leurs variations au cours du temps. Cependant, Kreiman et Sidtis (2011) soulignent la grande variété des attributs acoustiques pouvant être pris en compte. Il semble illusoire, selon ces auteures, d'en déterminer un nombre fixe. L'expérience perceptive de jugement de dissimilarité pour des paires de voix de Kreiman et Gerratt (1996) a corroboré cette pensée. L'objectif de l'étude était d'évaluer la fiabilité inter-juges, c'est-à-dire le degré d'accord entre les différents juges (Remacle et al., 2014). Cinq juges experts en perception vocale ont été recrutés, dont deux logopèdes, deux linguistes et un expert en linguistique et en logopédie. Un total de 632 paires de voix parmi 3160 ont été écoutées deux fois afin d'aboutir à un test-retest. La tâche consistait à évaluer, sur une échelle de 1 (identiques) à 7 (différentes), des paires de voix en termes de souffle et de raucité. La fiabilité inter-juges a été contrôlée par la présentation des mêmes paires de stimuli à tous les participants selon les mêmes conditions, dans un ordre aléatoire. Bien que les stimuli écoutés étaient identiques pour chaque participant, à aucun moment deux voix n'ont été regroupées dans une même catégorie. Il n'a pas été possible de dégager de dimensions communes sur lesquelles les participants se sont appuyés. L'emploi de stratégies perceptives très différentes a été constaté. Par conséquent, l'existence d'un espace perceptif commun a été remise en cause, réduisant la possibilité de construire une échelle standardisée pour l'évaluation de la qualité vocale (Kreiman & Gerratt, 1996). Cette qualité vocale serait plutôt le fruit d'une interaction entre le signal acoustique et l'auditeur (Kreiman & Sidtis, 2011). Autrement dit, les individus tendent à focaliser leur attention sur des aspects particuliers en fonction de leurs attentes perceptives (Kreiman et al., 1994). Alors que le modèle de la perception vocale de Belin et al. (2004) expose les étapes spécifiques de reconnaissance

de l'identité de l'individu à partir de sa voix (Belin et al., 2011), il paraîtrait que chaque individu explore ces chemins d'une manière qui lui est propre.

1.3. Voix prototypique

L'étude de Kreiman et Gerratt (1996) a mis à jour la variabilité interindividuelle concernant l'évaluation perceptive de la voix. Les stratégies perceptives diffèrent entre les individus et sont influencées par leurs attentes perceptives. Cependant, comment l'auditeur parvient-il à reconnaître une voix à partir de ses attentes perceptives et en quoi consistent ces dernières ?

Dépendante de la configuration de l'appareil phonatoire, la voix se compose de diverses caractéristiques physiques qui lui sont propres (Belin et al., 2004). Cette « signature vocale » (Belin et al., 2011, p.714) n'est toutefois le résultat que de légères fluctuations acoustiques. Une base acoustique commune serait partagée entre les différentes voix humaines (Belin et al., 2011). Il semblerait que pour reconnaître une voix, nous comparons les paramètres acoustiques de la voix entendue avec ceux d'une voix dite « **prototypique** » (Belin et al., 2011).
Les résultats de la tâche de reconnaissance vocale de Papcun et al. (1989) a mené à cette conclusion. Parmi un groupe de voix distractrices, 90 auditeurs devaient repérer la voix qu'ils avaient entendue, une, deux ou quatre semaines plus tôt. Celles-ci étaient réparties en deux catégories : les voix jugées faciles à retenir et les voix jugées difficiles à retenir. Afin de contrôler la fiabilité inter-juges, les participants ont été répartis aléatoirement dans deux groupes. Une corrélation significative a été observée entre les moyennes des deux groupes pour les scores de reconnaissance des voix. Les résultats montrent que les participants tendaient à reconnaître les voix distractrices « difficiles à retenir » comme la voix cible. Ceci n'a pas été observé pour les distracteurs « faciles à retenir ». Les voix « difficiles à retenir » seraient donc assimilées aux voix prototypiques. Au contraire, les caractéristiques acoustiques des voix « faciles à retenir » les éloigneraient de la prototypicalité. Papcun et al. (1989) ont conclu que plus la proximité entre la voix cible et la voix prototypique est grande, mieux la cible sera retenue. L'augmentation du délai entre la première et la seconde écoute engendrait un oubli des déviations des voix « faciles à retenir » par rapport à la prototypicalité. Ceci reflète l'instabilité de leurs représentations en mémoire. Les voix « faciles à retenir » deviennent de plus en plus similaires à celles des voix « difficiles à retenir ». Par conséquent, les auteurs suggèrent que le souvenir des individus est une voix définie par un prototype et des déviations autour de ce dernier. Ils ont proposé l'existence d'un modèle de mémoire à long terme fondé sur l'encodage

du stimulus vocal comme prototype ou membre central d'une catégorie, accompagné de diverses déviations possibles (Kreiman et al., 1990). Ces références prototypiques sont le résultat d'une exposition répétée à divers stimuli vocaux, induite par les expériences de vie (Papcun et al., 1989). Chaque prototype est propre à l'individu car il est le fruit de son vécu. Ce constat peut être lié avec l'observation de Kreiman et Gerratt (1996) selon laquelle les stratégies perceptives diffèrent entre les individus.

En 1991, Kreiman et Papcun ont mené une étude semblable auprès de 124 participants. Parmi eux, 24 ont été soumis à une tâche de discrimination de paires vocales et 100 ont réalisé une tâche de reconnaissance vocale. Au sein de l'épreuve de discrimination, les voix « faciles à retenir » étaient moins sujettes à des identifications erronées. La réponse étant donnée immédiatement, la perte mnésique de leurs caractéristiques distinctives était limitée. En ce qui concerne la tâche de reconnaissance vocale, soumise à un délai d'une semaine, les voix cibles « difficiles à retenir » tendaient à être correctement reconnues. Les distracteurs « difficiles à retenir » étaient, quant à eux, souvent faussement identifiés. Les analyses de Kreiman et Papcun (1991) concordent avec celles de Papcun et al. (1989). A nouveau, les résultats ont illustré une tendance à opter pour la voix sonnant « typiquement », à mesure que disparaissent les distinctions entre la voix et le prototype. Cette référence prototypique permet également de réduire la quantité d'énergie consommée pour traiter un signal vocal (Latinus et al., 2013).

1.4. Caractéristiques acoustiques vocales élémentaires

Afin de dépasser les fluctuations vocales interindividuelles, les auditeurs s'appuient sur des voix prototypiques. La prototypicalité interne s'organise autour de caractéristiques acoustiques précises qu'il semble intéressant de définir. Dans cette optique, plusieurs études ont cherché à déterminer les composantes acoustiques importantes de la voix dans le but de mieux comprendre l'interaction entre le signal vocal et la perception.

En 2010, Baumann et Belin ont cherché à identifier les attributs acoustiques cruciaux d'un point de vue perceptif. Ils ont recherché les indices sur lesquels se sont appuyés les 10 participants (5 hommes et 5 femmes de 23.9 ans en moyenne) pour discriminer 32 locuteurs (16 hommes et 16 femmes). Une tâche de jugement de similarité perceptive entre des paires de voyelles isolées a été employée. Cette recherche a mis en lumière l'existence d'un « **espace vocal acoustique** » (Latinus & Belin, 2011), au sein duquel les voix sont représentées individuellement par des points (voir figure 3). Plus la distance entre les points est importante, plus la différence perceptive entre les voix est nette (Latinus & Belin, 2011). Ainsi, une identité

vocale similaire sera associée à des voix dont la localisation est proche, alors que des identités vocales divergentes seront associées à des voix espacées les unes des autres (Latinus & Belin, 2011).

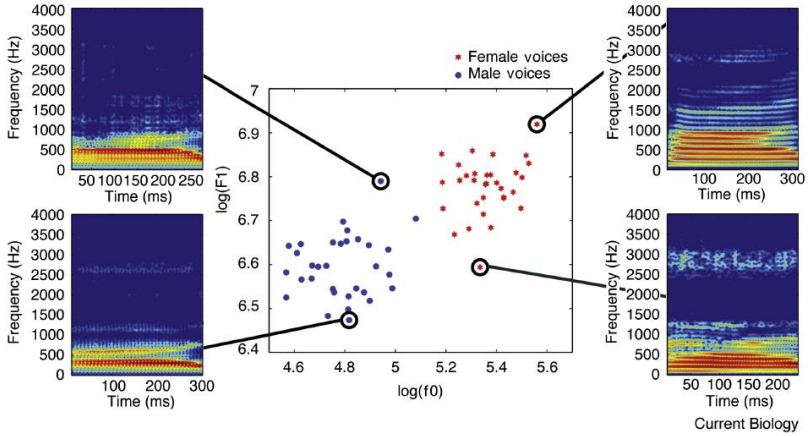

Figure 3 : Un espace vocal acoustique (Latinus & Belin, 2011)

Bien que de nombreuses dimensions acoustiques délimitent cet « espace vocal acoustique » (Latinus et al., 2013), Baumann et Belin (2010) ont démontré que deux d'entre elles étaient suffisantes pour percevoir une similarité vocale : la fréquence fondamentale (f_o) et les formants. Outre le fait qu'elles soient facilement mesurables, elles offrent une bonne représentation des voix féminines et masculines. La f_o illustre la contribution de l'étage laryngé. Elle correspond à la vitesse de répétition de la période (ou cycle) de l'onde acoustique (Johnson, 2003). Créée par le mécanisme d'ouverture et de fermeture des plis vocaux dans le larynx (Belin et al., 2004), elle est déterminée par leur longueur et leur masse (Fitch, 1997). Les formants reflètent l'implication de l'étage supra-laryngé. Ils sont des bandes de fréquences amplifiées qui dépendent de la géométrie du tractus vocal, c'est-à-dire de sa taille et de sa forme (Fitch, 1997). En accord avec l'étude menée par Baumann et Belin (2010), Latinus et al. (2013) ont exposé la f_o et la dispersion des formants comme dimensions acoustiques importantes de l' « espace vocal acoustique ». La dispersion des formants se définit par la « différence moyenne entre les fréquences de formants qui se suivent » (Fitch, 1997, p.1213). Un troisième paramètre peut, selon eux, compléter la délimitation de cet espace et, par conséquent, la définition des voix prototypiques au sein de ce dernier. Il s'agit du rapport harmoniques/bruit. Il correspond au rapport entre la quantité d'énergie périodique (les harmoniques), produite par la vibration des plis vocaux, et la quantité d'énergie apériodique (le bruit), causée par l'interaction du flux d'air avec les parois et obstacles du tractus vocal, dans le signal de parole (Hillenbrand, 1987). Notre perception de la voix semble dès lors être organisée autour d'attributs acoustiques moyennés : la fréquence fondamentale, la dispersion des formants et le

rapport harmoniques/bruit. L'identité vocale est définie par une comparaison avec des normes, à partir desquelles un prototype vocal féminin et masculin sont construits. La distance par rapport à la moyenne établie paraît donc être un indice pertinent pour la distinction des voix (Latinus et al., 2013). Ces auteurs soulignent cependant le fait que la complexité de l' « espace vocal acoustique » s'étend au-delà des trois dimensions citées mais que ces dernières fournissent une représentation proche de la réalité.

En résumé, le lobe temporal supérieur postérieur bilatéral est une région cérébrale critique pour la perception et le traitement de la parole et de la voix. Afin d'analyser un signal vocal, nous évaluons sa qualité vocale, qui se caractérise par une importante variabilité interindividuelle. L'évaluation de la qualité vocale est influencée par les attentes perceptives de l'auditeur. Ces attentes perceptives s'organisent autour de voix prototypiques, encodées dans un « espace vocal acoustique » tridimensionnel.

2. Hautes fréquences

2.1. Une longue ignorance

Les hautes fréquences (HF) correspondent aux fréquences dont l'énergie acoustique se situe au-delà de 5 kHz (Monson, Hunter et al., 2014). Cette limite fréquentielle ne fait pas consensus et reste propre au domaine de la parole humaine (R. Blandin, communication personnelle, 6 mai 2020). Boyd-Pratt et Donai (2020) distinguent le terme « hautes fréquences » (> 5 kHz) du terme « hautes fréquences étendues » (HFE) (> 6-8 kHz). Dans ce mémoire, nous appelons « hautes fréquences » toutes les fréquences supérieures à 5 kHz, c'est-à-dire à la fois les HF et les HFE. La figure 4 situe les HF dans un spectre typique de la parole (Monson, Hunter et al., 2014).

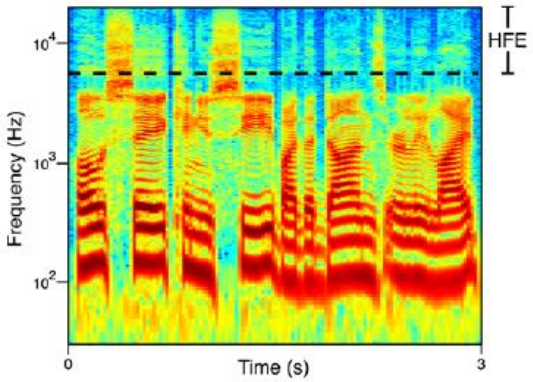

Figure 4 : Spectrogramme de la parole à large bande représentant la phrase « *Oh say can you see by the dawn's early light* » produite par un homme (Monson, Lotto et al., 2014)

2.1.1. Un paramètre acoustique peu important

Les humains sont capables de percevoir un large registre fréquentiel qui s'étend de 0.02 à 20 kHz (Monson & Caravello, 2019). Pourtant, les fréquences supérieures à 5 kHz ont été dépeintes avec une importance mineure pour la perception de la parole. La principale raison est que l'énergie en basse fréquence est présentée comme suffisante pour l'intelligibilité de la parole (Monson, Hunter et al., 2014). Motivées par le développement du téléphone, les recherches avaient l'objectif de déterminer la bande passante minimale nécessaire à la diffusion d'informations intelligibles (Monson, Hunter et al., 2014).

Durant la première moitié du 20ème siècle, le modèle théorique intitulé « index articulatoire », récemment renommé « index d'intelligibilité de la parole », a été créé (Boyd-Pratt & Donai, 2020 ; Monson, Hunter et al., 2014). Les mesures de ce modèle ont été effectuées dans un environnement calme, auprès d'adultes sans antécédent auditif. Les résultats ont montré un taux d'intelligibilité de 95% pour des monosyllabes de type consonne-voyelle-consonne (CVC) lorsqu'un filtre passe-bas à 4 kHz était utilisé (French & Steinberg, 1947, cités par Monson, Hunter et al., 2014 ; Fletcher & Galt, 1950, cités par Monson, Hunter et al., 2014). Le filtre passe-bas permet de ne transmettre que les basses fréquences contenues dans le signal acoustique à partir d'un seuil déterminé. Les fréquences inférieures à 4 kHz permettaient donc de conserver une très bonne intelligibilité pour le signal de parole. Un taux d'intelligibilité maximal était acquis pour les mêmes stimuli avec un filtre passe-bas à 7 kHz (French & Steinberg, 1947, cités par Monson, Hunter et al., 2014 ; Fletcher & Galt, 1950, cités par Monson, Hunter et al., 2014). Ces observations ont soumis l'idée de la faible utilité des HF pour rendre la parole intelligible dans un environnement calme. En 2003, Johnson affirme que les humains exploitent majoritairement l'information transmise par les fréquences inférieures à 2 kHz lors du traitement de la parole. Avec une bande passante limitée à 4.2 kHz, Englert (1986) a obtenu 97% d'intelligibilité pour des monosyllabes de type consonne-voyelle (CV) et voyelle-consonne (VC). La tâche perceptive a eu lieu dans une pièce calme, auprès de 15 adultes sans antécédent auditif. Chacun était testé individuellement. Les résultats suggèrent l'importance de l'énergie en basse fréquence. Néanmoins, l'auteure souligne qu'une parfaite intelligibilité n'a pas été atteinte. Il semblerait que la bande passante limitée aux basses fréquences ne fournisse pas la totalité des informations nécessaires à la compréhension du signal.

2.1.2. Une faible puissance technologique

Des contraintes technologiques et informatiques ont entravé l'étude des HF (Monson, Hunter et al., 2014). Les limitations en puissance et rapidité des transducteurs et des équipements électroniques ont empêché d'aboutir à leur représentation fidèle (Monson, Hunter et al., 2014 ; Monson & Caravello, 2019). Leur transcription était pauvre dans le tracé du spectre par les appareils d'analyses, en raison de leur faible performance en HF. Malgré les avancées dans le domaine informatique dans les années 1970 (Kuligowska et al., 2018), la tendance à poursuivre les recherches sans prendre en compte les HF s'est maintenue et est restée dominante, bien qu'aucune étude ne soit parvenue à démontrer leur inutilité perceptive (Monson, Hunter et al., 2014).

2.1.3. Un paramètre acoustique difficile à maîtriser

Des difficultés supplémentaires liées aux aspects acoustiques des HF ne sont pas à négliger. Les HF se caractérisent par leur grande complexité et leur variabilité dans le temps et l'espace. Contrairement à la stabilité de l'énergie en basse fréquence, elles présentent un spectre bien plus hétérogène. L'énergie acoustique dans les HF est moins intense que dans les basses fréquences dans la parole humaine (Monson et al., 2012). Les auteurs ont montré une diminution progressive d'environ 20 à 40 dB SPL entre 0 et 5 kHz. De plus, la longueur d'onde des HF, c'est-à-dire la distance parcourue par l'onde durant une oscillation dans l'environnement, est plus courte que celle des basses fréquences. Ceci rend difficile leur modélisation physique. Cette complexité est accentuée par la forme irrégulière du tractus vocal durant la production de parole (Monson, Hunter et al., 2014). Afin d'offrir une représentation acoustique la plus précise possible, il est nécessaire d'avoir recours à du matériel d'enregistrement audio de haute qualité, notamment concernant les microphones et les cartes son (Monson, Hunter et al., 2014). Les auteurs ajoutent que la propagation des HF est plus sensible aux effets de diffusion et de directivité. L'effet de diffusion correspond à la réflexion d'une onde acoustique dans de multiple directions tandis que l'effet de directivité correspond à la variation d'amplitude avec la direction. Par exemple, les HF d'un signal de parole produit par un locuteur ont une plus grande amplitude en face de lui que derrière lui (R. Blandin, communication personnelle, 22 juin 2020). Ces données impliquent une certaine connaissance des attributs acoustiques du cadre expérimental. A titre illustratif, il faut envisager les caractéristiques réflexives des surfaces murales. Une plus grande vigilance par rapport à la position des haut-parleurs et des microphones est donc requise pour l'étude des HF.

2.2. Un regain intérêt

2.2.1. Perception des HF dans la parole humaine

L'intérêt pour la gamme fréquentielle haute du spectre vocal a connu un nouvel essor durant la dernière décennie (Boyd-Pratt & Donai, 2020). Plusieurs études ont cherché à démontrer la sensibilité perceptive du système auditif humain à la présence de HF au sein de la parole. L'étude de Monson et al. (2011) a été menée auprès de 30 auditeurs (16 hommes et 14 femmes) d'une moyenne d'âge de 29.5 ans. Tous ont subi une audiométrie afin de s'assurer que leurs seuils auditifs binauraux étaient inférieurs ou égaux à 15 dB HL pour l'ensemble des fréquences testées (de 0.25 à 8 kHz). L'expérience a eu lieu dans une pièce insonorisée. Elle consistait en une épreuve de détection de l'atténuation de l'énergie des HF dans une voyelle [a] isolée, chantée sans vibrato à différents niveaux d'intensité. Plus précisément, l'énergie des bandes fréquentielles de 8 et 16 kHz était atténuée de façon individuelle. Les résultats ont montré la capacité des auditeurs à détecter des changements d'intensité dans les HF, même lorsqu'ils étaient légers. De meilleures performances ont été décelées pour la bande fréquentielle de 8 kHz, suggérant sa plus forte contribution dans la perception de parole. Plus de la moitié des participants parvenaient également à percevoir les modifications à 16 kHz. Cette recherche a mis en lumière la sensibilité du système auditif humain aux HF dans la parole. D'après Monson et al. (2011), ces résultats suggèrent que les HF contiennent de l'information pertinente dans le signal de parole.

Plus récemment, Monson et Caravello (2019) ont cherché à définir la fréquence maximale du filtre passe-bas à partir de laquelle les 21 participants (7 hommes et 14 femmes), âgés de 19 à 27 ans, n'étaient plus capables de percevoir les HF dans la parole. Leurs seuils auditifs pour les fréquences de 0.5 à 16 kHz ont été mesurés avec une audiométrie. Les participants rapportant des antécédents auditifs et ne présentant pas un seuil auditif inférieur ou égal à 20 dB HL à toutes les fréquences testées pour au moins une oreille étaient exclus. Une tâche de détection de la différence a été proposée, entre un signal de référence comportant un filtre passe-bas à 22 kHz et un signal test présentant un filtre passe-bas à des fréquences progressivement plus graves. Le signal était la phrase « amend the slower page », enregistrée par deux hommes et deux femmes. Elle a été choisie car elle comporte au moins une fois chaque trait phonétique distinctif (voisement, nasalité, etc.). Avec une limite moyenne respectivement située à 13.1 kHz et 12.8 kHz pour la parole de femmes et d'hommes, la capacité des participants à percevoir une différence dans les HF de la parole jusqu'à 13.1 kHz a été mise à jour. Cette faculté de détection semble indiquer que les HF jusqu'à 13 kHz environ peuvent être porteuses d'informations

pertinentes pour le signal de parole (Monson, Hunter et al., 2014). L'utilité des HF dans la perception de la parole est abordée dans la partie 2.3..

La population testée des études susmentionnées est limitée à des jeunes adultes. Ceci pose la question de la généralisation à toutes les tranches d'âge de la capacité à percevoir les HF dans la parole humaine. De plus, les résultats diffèrent légèrement selon que le signal de parole ait été produit par une femme ou par un homme (Monson, Hunter et al., 2014). Enfin, bien que le stimulus dans l'étude de Monson, Hunter et al. (2014) ait été une phrase phonétiquement représentative des différents traits distinctifs entre phonèmes, les auteurs n'ont pas exploré le lien potentiel entre le phonème et sa composition acoustique en HF. Les effets de l'âge de l'auditeur, du genre du locuteur et du type de matériel linguistique sur les HF dans la parole humaine sont abordés dans les points 3.2.2.1, 3.2.2.2. et 3.2.2.3..

2.2.2.1. Effet de l'âge de l'auditeur sur la perception des HF dans la parole humaine

Contrairement aux adultes, les enfants et les adolescents présentent une grande sensibilité aux HF dans le signal de parole (Monson & Caravello, 2019 ; Rodríguez Valiente et al., 2014). Cette faculté se marque toutefois par un déclin progressif depuis l'enfance jusqu'à l'âge adulte (Rodríguez Valiente et al., 2014). L'avancée en âge engendre un vieillissement du système auditif humain, caractérisé par une augmentation des seuils auditifs et une diminution des otoémissions acoustiques[2] (OEA) (Gordon-Salant, 2005 ; Hunter et al., 2020). Alors que la recherche d'une éventuelle perte auditive se fait généralement vers 60-70 ans, Hunter et al. (2020) ont montré qu'une modification plus précoce de la sensibilité auditive est observable. Une évaluation des seuils auditifs et des OEA dans les HF a été réalisée auprès 313 participants (118 hommes et 195 femmes) âgés de 10 à 65 ans. Ces derniers présentaient un seuil auditif inférieur ou égal à 20 dB HL jusqu'à 4 kHz, un fonctionnement de l'oreille moyenne normal et n'avaient aucun antécédent auditif. Les résultats ont mis en lumière un déclin systématique des seuils auditifs et des OEA avec l'âge, particulièrement marqué dans les HF. Cette perte de sensibilité était majeure à 12.5 kHz pour les OEA et à 16 kHz pour les seuils auditifs chez les participants de 36 à 45 ans. Le déclin de ces deux paramètres tend à se poursuivre dans une gamme de fréquences plus étendue, à mesure que l'individu vieillit (Hunter et al., 2020). L'âge exerce donc une influence non négligeable sur la faculté de perception des HF. Les

[2] « Sons émis par [les cellules ciliées externes de] la cochlée en l'absence (OEA spontanées) ou en réponse à une stimulation auditive [...] (OEA provoquées) » (Collet, 1994, p.45)

changements auditifs précoces soulignent l'enjeu d'un diagnostic anticipé à l'aide d'une extension du registre fréquentiel testé lors de l'audiométrie.

2.2.2.2. Effet du genre du locuteur sur la perception des HF dans la parole humaine

La quantité d'énergie dans les HF diffère entre les voix féminines et les voix masculines en fonction des bandes d'octave (voir figure 5) (Monson et al., 2012). L'étude de Stelmachowicz et al. (2001) s'est appuyée sur une tâche de perception de syllabes CV et VC contenant chacune une consonne fricative ([s], [f] ou [θ]) et la voyelle [i]. Un groupe de 40 participants avec une audition normale et un second de 40 participants avec un trouble auditif ont été créés. Chacun était composé de 20 adultes et 20 enfants. La consigne était de cliquer sur le graphème « s », « f » ou « th » sur l'écran d'ordinateur pour indiquer la syllabe entendue. Les résultats ont montré que la performance de reconnaissance maximale différait pour tous les participants selon que le stimulus ait été produit par un homme, une femme ou un enfant. Concernant le stimulus masculin, la reconnaissance syllabique était meilleure quand la bande passante possédait un filtre passe-bas à 4-5 kHz. Pour les signaux de parole produits par les femmes et les enfants, le score maximal était obtenu lorsque le filtre passe-bas était situé à 9 kHz. Ces résultats ont été corroborés par l'étude de Monson et al. (2012), dans laquelle des effets de genre ont été démontrés. L'analyse de 20 phrases produites par 7 hommes et 8 femmes a montré que l'énergie en HF était plus importante dans les HF basses (au niveau de la bande d'octaves de 6.3 kHz) pour les voix masculines. En revanche, à mesure que la fréquence augmente, la quantité d'énergie en HF dans les voix féminines tend à augmenter, jusqu'à dépasser celle des hommes. Plus précisément, elle était supérieure pour 5 des 6 bandes d'octave et de tiers d'octave testées (8 kHz ; 10.1 kHz ; 12.7 kHz ; 16 kHz ; 20.2 kHz).

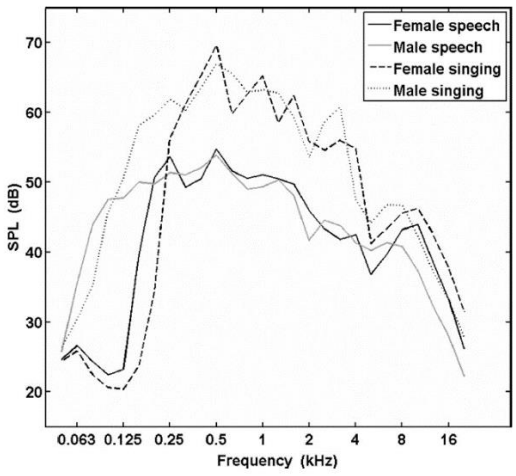

Figure 5 : Répartition de l'énergie acoustique dans les HF dans de la parole et du chant en fonction du genre du locuteur (Monson et al., 2012)

2.2.2.3. Effet du type de matériel linguistique sur la perception des HF dans la parole humaine

D'après Monson et Buss (2019), une grande partie des indices acoustiques de la parole se limite aux basses fréquences. Tel est le cas pour les formants des voyelles. Au contraire, les consonnes se caractérisent par une importante énergie dans les HF, au-delà de 5 kHz (Monson, 2011). C'est particulièrement le cas pour les fricatives non voisées [s], [ʃ], [f] et [θ] (Monson, Hunter et al., 2014). Monson et al. (2012) les ont enregistrées auprès de 15 locuteurs sains (7 hommes et 8 femmes) âgés en moyenne de 28.5 ans. Chacun était expert dans l'utilisation de sa voix grâce à au moins deux ans de formation vocale en chant après l'école secondaire. L'analyse acoustique des fricatives montre un effet du lieu d'articulation sur la répartition de l'énergie des HF dans le spectre. La figure 6 présente les résultats obtenus. Les fricatives [s] et [ʃ] se distinguent des fricatives [f] et [θ], qui présentent une quantité d'énergie dans les HF moins importante. La fricative [f] comporte davantage de HF dans la bande d'octaves de 8 kHz. La quantité d'énergie dans les HF diminue ensuite légèrement. L'inverse est observé avec la fricative [θ] qui présente plus d'énergie dans les HF entre 8 et 16 kHz.

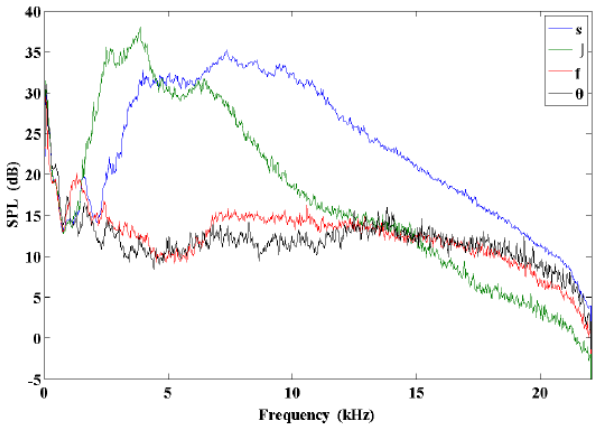

Figure 6 : Spectre acoustique à long terme moyenné pour chaque fricative non voisée ([s], [ʃ], [f], [θ]) produite par des locuteurs sains (Monson et al., 2012)

L'approfondissement des analyses a mis à jour un effet d'interaction entre le genre et la quantité d'énergie dans les HF au sein de certaines fricatives et de certaines bandes d'octave (Monson et al., 2012). Les fricatives [s] et [ʃ] comportent une plus grande quantité d'énergie dans les HF dans la parole de femmes. Concernant le [f], les HF se concentrent principalement dans la bande d'octave de 8 kHz pour la parole d'hommes et dans la bande d'octave de 16 kHz pour la parole de femmes.

En 1995, Shadle et Scully ont indiqué que les pics spectraux du [s] pouvaient dépendre du contexte. En effet, l'arrondissement des lèvres lors de l'articulation de ce phonème a contribué

à une déviation du pic spectral depuis 5.5 kHz pour les non-mots « pasa » et « pisi » à 7.5 kHz pour le non-mot « pusu ». Monson (2011) a suggéré la possibilité d'une modification volontaire de l'énergie en HF par la modulation de la géométrie du tractus vocal.

2.3. Rôles des hautes fréquences

D'après Monson, Hunter et al. (2014), les usages des HF dans la parole humaine seraient multiples. Leur pertinence perceptive a été traduite par diverses études qui ont exposé leur contribution pour la localisation de la source sonore, la reconnaissance du locuteur, l'intelligibilité de la parole et la qualité du signal de parole en termes d'aspect naturel (Best et al., 2005 ; Monson et al., 2019 ; Monson, Hunter et al., 2014). Dans ce mémoire, seuls les rôles liés à l'intelligibilité et à l'aspect naturel de la parole sont détaillés.

2.3.1. Intelligibilité de la parole

Plusieurs études ont souligné l'utilité des HF pour l'intelligibilité de la parole. Lippmann (1996) a montré que l'énergie acoustique située dans les HF permettait la redondance de l'information. Administrée à 3 participants sans trouble auditif, la tâche consistait à transcrire la voyelle et les consonnes perçues dans des syllabes CVC produites par une femme. Un filtre passe-haut a été mis en place à 3.15, 4, 5, 6.3, 8 puis 10 kHz. Il a permis de ne transmettre que les fréquences supérieures aux seuils mentionnés. Un bruit blanc, filtré à 10 kHz par un filtre passe-bas, a été ajouté pour s'assurer de l'absence de signal résiduel en dehors des bandes passantes. Lorsque la bande fréquentielle moyenne (entre 0.8 et 4 kHz) était indisponible en raison du filtre passe-haut, les auditeurs s'appuyaient sur les fréquences supérieures. L'intelligibilité des consonnes étant conservée, l'auteur a suggéré que l'énergie contenue au-delà de 8 kHz comporte des indices significatifs pour leur identification et discrimination. Cependant, le faible nombre de participants limite la généralisation des résultats. A l'aide d'une tâche de catégorisation de voyelles et de consonnes, Vitela et al. (2015) ont obtenu des résultats similaires auprès d'une plus grande cohorte de 30 jeunes adultes, sans antécédent auditif. Les stimuli consistaient en des syllabes CV produites par un homme et une femme. Un filtre passe-haut à 5.7 kHz a été ajouté avec un bruit de masque dans les basses fréquences. La performance des sujets dépassait le hasard bien que les basses fréquences aient été retirées du signal. Ces deux recherches ont conclu que les HF contribuent à la perception de la parole par leur apport d'indices phonétiques. La présence des HF est particulièrement importante pour assurer l'intelligibilité des consonnes. Ceci s'illustre à travers l'exemple du téléphone, pour lequel la bande passante s'étend uniquement de 0.3 à 3.4 kHz (Monson, Hunter et al., 2014). L'absence

de transmission des HF entrave l'intelligibilité des consonnes. A titre illustratif, il n'est pas rare de pallier ce manque d'intelligibilité par la répétition de la consonne initiale pour les prénoms : « F comme Fanny, S comme Sophie... ».

Les HF semblent jouer un rôle particulier lorsque l'environnement est bruyant. L'identification de consonnes dans le bruit paraît facilitée quand la bande passante est haute (de 3.5 à 10 kHz), plutôt que basse (de 0 à 3.5 kHz) (Apoux & Bacon, 2004). Cette donnée suggère que l'effet Lombard a un impact positif sur la perception de la parole dans un environnement bruyant. Automatiquement mis en place par les locuteurs face au bruit environnant, il consiste à hyper-articuler en augmentant l'intensité vocale, la fréquence fondamentale et la durée des voyelles et des consonnes (Castellanos et al., 1996 ; Garnier & Henrich, 2014 ; Garnier et al., 2018). La parole gagne alors en intelligibilité (Castellanos et al., 1996). Zadeh et al. (2019) se sont aussi penchés sur le rôle des HF pour percevoir la parole dans un environnement bruyant. Leur étude consistait en une écoute de chiffres enregistrés et masqués par du bruit à l'aide d'un filtre passe-bas à 2, 4 puis 8 kHz. Du bruit à large bande était ajouté pour simuler un contexte de conversation. Les résultats ont montré que les 60 participants (21 hommes et 39 femmes de 29.6 ans en moyenne) ayant rapporté des difficultés d'écoute dans le bruit présentaient une perte auditive dans les HF. En outre, les chiffres non filtrés étaient plus intelligibles par rapport aux chiffres filtrés à 8 kHz. Les auteurs ont conclu que les HF supérieures à 8 kHz sont susceptibles de significativement contribuer à l'intelligibilité de la parole par la transmission d'indices pertinents. Certaines données obtenues par Badri et al. (2011) semblent aussi montrer qu'une perception dégradée des HF est corrélée aux difficultés de perception de la parole dans le bruit. Ces différentes études s'accordent sur l'importance des HF pour l'intelligibilité du signal de parole, notamment dans un environnement bruyant.

2.3.2. Aspect naturel de la parole

Il semblerait que les HF jouent également un rôle pour l'aspect naturel de la parole. En 2003, Moore et Tan ont administré une tâche de jugement subjectif de la qualité vocale de deux phrases générées par la synthèse par concaténation. L'une était produite avec une voix masculine et l'autre avec une voix féminine. Les seuils auditifs des 10 auditeurs âgés de 15 à 31 ans a été vérifiée avec une audiométrie. La tâche consistait à évaluer la qualité du stimulus sur une échelle de 1 (pas du tout naturel) à 10 (très naturel). Les résultats ont montré l'importance de la présence des HF dans la perception du signal. Le filtre passe-bas à 10.9 kHz rendait la parole plus naturelle, contrairement au filtre passe-bas à 7 kHz. En revanche, aucune différence perceptive n'a été relevée lorsque les bandes fréquentielles limitées à 10.9 et 16.9

kHz ont été comparées. Cette préférence pour un signal possédant des fréquences élargies est confirmée par la recherche de Füllgrabe et al. (2010). Les stimuli étaient des extraits de 2 à 4 secondes tirés d'un discours, de mots isolés ou de syllabes voyelle-consonne-voyelle (VCV). Leurs résultats démontrent que les stimuli étaient perçus comme plus clairs et agréables, par les 4 participants normo-entendants (1 homme et 3 femmes de 21 à 38 ans), quand la bande fréquentielle s'étendait jusqu'à 10 kHz. Les signaux limités en-deçà de 7.5 kHz paraissaient moins appréciés. Bien que ces études illustrent la contribution des HF dans l'aspect naturel du signal de parole, leurs résultats restent à interpréter avec prudence en raison du nombre limité de participants. Etudier la contribution des HF pour l'aspect naturel de la parole auprès d'une cohorte plus importante semble donc intéressant pour généraliser les résultats. Monson, Hunter et al. (2014) précisent que l'insertion de HF dans la synthèse de la parole, couplée à une modélisation du tractus vocal, serait susceptible d'améliorer la composante naturelle du signal de parole artificielle.

En résumé, les HF ont longtemps été ignorées dans les études sur la perception de la parole, en raison de l'importance perceptive mineure qui leur a été associée, de leurs difficultés de maîtrise acoustique et d'une faiblesse technologique. Pourtant, le système auditif humain est capable de percevoir les HF dans la parole humaine. Divers facteurs influencent cette capacité, comme l'âge de l'auditeur, son genre et la composition phonétique du signal de parole. La faculté de perception des HF dans la parole reflète leur utilité perceptive, notamment pour l'intelligibilité et l'aspect naturel du signal de parole. Comprendre et mieux définir ce paramètre acoustique offrent de multiples perspectives. En particulier, ces nouvelles connaissances sont susceptibles d'enrichir le champ de la synthèse de la parole. Une amélioration de la flexibilité et de la qualité de la voix de synthèse est rendue possible par la définition plus fine des différents marqueurs acoustiques et de leurs enjeux (Kreiman & Sidtis, 2011).

3. Méthodes de synthèse de la parole

Probablement l'outil de communication le plus sollicité par les humains (Holmes & Holmes, 2001), la parole est un élément prégnant de notre quotidien. Une volonté de transposer cette capacité humaine aux machines existe depuis longtemps. Grâce à l'avènement des technologies électroniques, il est possible de répondre à cette volonté par le recours à la synthèse de la parole (Holmes & Holmes, 2001 ; Huang, 2001 ; Shroeter, 2005). Elle consiste à produire artificiellement la parole humaine (Lukose & Upadhya, 2017).

3.1. Bref historique des méthodes de synthèse de la parole

Des tentatives de synthèse de la parole, plus ou moins réussies et datant parfois du 18ème siècle, peuvent être rapportées (Huang, 2001). Un intérêt grandissant s'est développé pour le traitement de l'information par les machines au fil du temps (Holmes & Holmes, 2001 ; Huang, 2001 ; Shroeter, 2005). L'utilisation d'ordinateurs plus puissants et de logiciels plus performants a offert de nouvelles perspectives quant à la synthèse de la parole (Huang, 2001 ; Shroeter, 2005). Avec l'objectif d'obtenir un signal de parole idéal, elle vise à optimiser deux attributs particuliers : l'intelligibilité et l'aspect naturel (Tabet & Boughazi, 2011). L'intelligibilité réfère au degré à partir duquel le signal de parole est compris (Anand & Stepp, 2015). Elle dépend de « l'intégration et la coordination de tous les autres processus impliqués dans la parole, en particulier la respiration, la phonation et l'articulation » (Yorkson & Beukelman, 1981, cités par Robertson & Thomson, 1999). L'aspect naturel de la parole vise à réduire la distance perceptive entre la parole synthétique et la parole humaine (Tabet & Boughazi, 2011).

L'un des systèmes utilisés est le système « text-to-speech ». Il s'agit d'une méthode de synthèse automatique de la parole, pour laquelle deux principales étapes sont requises pour aboutir à la lecture à voix haute de n'importe quel support (Dutoit, 1997 ; Kuligowska et al., 2018). Ces étapes réfèrent à l'analyse du texte à travers une description de ses règles linguistiques, suivie de la synthèse de la parole par une production des sons de parole à partir des règles linguistiques décrites (Kuligowska et al., 2018). Durant la première étape, une conversion grapho-phonémique est appliquée : les graphèmes (lettres) sont convertis en phonèmes (sons) (Dutoit, 1997). Ils sont ensuite fusionnés et diffusés dans un haut-parleur. Le texte entré dans le logiciel peut alors être écouté. Les annonces automatiques dans les gares, l'aide à la navigation routière dans les GPS, les logiciels de traduction (Google Translate, DeepL, etc.) sont autant d'applications possibles du système « text-to-speech » (R. Blandin, communication personnelle, 6 août 2021).

Les sons de parole peuvent être produits avec différentes méthodes. Dans le cadre de ce mémoire, seules trois d'entre elles sont détaillées : la synthèse par formants (ou par règles), la synthèse par concaténation et la synthèse articulatoire (Khan & Chitode, 2016). La dernière méthode étant celle employée dans notre recherche, elle est davantage approfondie.

3.1.1. Synthèse par formants (ou synthèse par règles)

A son apogée entre les années 1960 et 1980 (Dutoit et al., 2002), la synthèse par formants (ou par règles) se fonde sur une démarche paramétrique. Par l'application de règles, les formants sont modélisés par un contrôle de leurs fréquences, de leurs amplitudes et des caractéristiques des plis vocaux (Huang, 2001). Elle s'appuie sur le modèle source-filtre, dans lequel la source correspond à la mise en vibration des plis vocaux et le filtre réfère au tractus vocal (Khan & Chitode, 2016 ; Lukose & Upadhya, 2017). La synthèse par formants permet de générer un signal à partir de sources et de résonateurs suite au « dessin » du spectrogramme par le phonéticien. Un spectrogramme est une représentation visuelle tridimensionnelle d'un signal vocal et/ou de parole, qui décompose le signal en termes de temps, de fréquence et d'amplitude (voir figure 4, page 11) (Sicard & Menin-Sicard, 2021). La synthèse par formants s'appuie sur l'expertise du chercheur qui lui permet de décoder le spectrogramme. Il devrait être capable de retranscrire ce qu'il observe au sein d'un spectrogramme artificiel pour une combinaison de phonèmes donnée (Dutoit et al., 2002). Cette méthode permet aussi l'ajout de certaines dimensions telles que la nasalité, des antirésonances ou encore des effets de rayonnement aux lèvres et aux narines (Huang, 2001). Ne dépendant pas d'une base de données d'échantillons de parole, les coûts de mémoire et de traitement informatiques s'en trouvent réduits. La synthèse par formants a l'avantage de fournir un signal très intelligible, et ce, même à des débits rapides (Tabet & Boughazi, 2011).

Plusieurs limites peuvent cependant être relevées. Les formants étant continuellement influencés par le contexte articulatoire et prosodique, les règles à appliquer sont fortement complexifiées (Huang, 2001). De nombreux essais et erreurs sont nécessaires avant d'aboutir à un signal satisfaisant. Beaucoup de temps et d'efforts sont requis pour développer ce système (Dutoit, 1997). La parole humaine est aussi réputée pour sa richesse et sa complexité. Une large gamme de paramètres est à prendre en compte lors de la synthèse (Huang, 2001). De plus, il s'agit d'un système dépendant de l'expertise humaine (Dutoit et al., 2002). Bien que le signal soit très intelligible, il reste artificiel, robotique, limitant sévèrement son aspect naturel (Dutoit et al., 2002 ; Huang, 2001 ; Lukose & Upadhya, 2017 ; Tabet & Boughazi, 2011).

3.1.2. Synthèse par concaténation

La synthèse par concaténation est la méthode la plus utilisée dans les applications commerciales (Khan & Chitode, 2016). La concaténation consiste en la juxtaposition de segments acoustiques coarticulés. Ils prennent la forme de phonèmes, de diphones, de triphones,

de syllabes ou de mots, tous préenregistrés dans des bases de données (Dutoit et al., 2002 ; Tabet & Boughazi, 2011). Deux méthodes différentes de synthèse par concaténation sont présentées, à savoir la synthèse par diphones et la synthèse par sélection d'unités.

La synthèse par diphones s'appuie sur un inventaire de diphones (Tabet & Boughazi, 2011). Défini par Dutoit, un diphone est « une unité acoustique qui commence au milieu de la zone stable d'un phonème et se termine au milieu de la zone stable du phonème suivant » (2002, p.3). Son caractère transitif offre l'atout majeur d'inclure la notion de coarticulation (Tabet & Boughazi, 2011). Souvent, la phase de concaténation est suivie d'une étape de réduction des discontinuités contextuelles (Dutoit et al., 2002). Les diphones juxtaposés ne s'enchaînent pas toujours avec limpidité au niveau des points de concaténation. Le recours au modèle « harmonic plus noise » (HNM) permet de remédier à ce problème (Stylianou, 2001). Il suggère que le signal de parole est constitué de deux parties : des harmoniques et du bruit (Tabet & Boughazi, 2011). Les harmoniques sont des multiples entiers de la fréquence fondamentale (Castellengo, 2015). Ils renvoient aux éléments périodiques, c'est-à-dire aux éléments qui reviennent à intervalles réguliers (Castellengo, 2015). Le bruit réfère aux éléments apériodiques, dont la survenue est aléatoire. Ce modèle permet de modifier certains paramètres acoustiques de l'onde, tels que la fréquence fondamentale, la fréquence vocale maximale (le nombre d'harmoniques) et l'enveloppe spectrale (les amplitudes et les phases harmoniques). Une minimisation des incompatibilités résiduelles entre diphones est engendrée par l'application de l'HNM (Dutoit et al., 2002 ; Huang, 2001). La synthèse par concaténation par diphones offre un signal intelligible (Dutoit et al., 2002). Toutefois, une sonorité robotique persiste (Lukose & Upadhya, 2017).

La synthèse par sélection d'unités est une autre méthode de synthèse par concaténation. Il s'agit d'un procédé de modelage statistique à partir de segments de parole plus larges que celle de la synthèse par diphones (Lukose & Upadhya, 2017). Ce procédé permet la sélection dynamique des segments acoustiques les plus appropriés en termes prosodiques (Huang, 2001 ; Khan & Chitode, 2016). A nouveau, l'un des objectifs est de réduire les discontinuités acoustiques en optant pour les unités présentant des extrémités spectrales proches (Dutoit et al., 2002). Divers contextes articulatoires et prosodiques sont illustrés par une multitude d'exemples pour chaque segment (Huang, 2001 ; Khan & Chitode, 2016). La sélection d'unités acoustiques est rendue automatique par l'application d'un algorithme (Tabet & Boughazi, 2011). Ce dernier permet de minimiser les coûts de représentation de la parole et de concaténation (Dutoit et al., 2002). La double prise en compte des aspects prosodiques et

concaténatifs par le système permet la production d'un signal bien plus naturel que la synthèse par diphones (Khan & Chitode, 2016 ; Lukose & Upadhya, 2017 ; Tabet & Boughazi, 2011). Une confusion avec une parole humaine est d'ailleurs parfois possible (Dutoit et al., 2002 ; Lukose & Upadhya, 2017). Ces divers aspects font de la synthèse par sélection d'unités la méthode de synthèse la plus utilisée dans les applications commerciales (Khan & Chitode, 2016).

La synthèse par concaténation comporte toutefois certaines limites. Un premier problème concerne la sélection des unités acoustiques (Huang, 2001). Chaque langue possède ses propres phonèmes et diphones. Le répertoire proposé ne peut donc pas faire preuve d'exhaustivité, ce qui restreint le nombre d'effets articulatoires présents dans les bases de données. Il en résulte un manque de flexibilité (Huang, 2001). De plus, la parole apparaît souvent hyper articulée, entravant l'aspect naturel de l'intonation (Dutoit et al., 2002). Plus les unités acoustiques sont longues, moins il y a de points de concaténation et plus le signal semble naturel. Ces unités nécessitent d'être stockées, augmentant le coût de mémoire informatique (Tabet & Boughazi, 2011). Des segments plus courts permettent de limiter ce coût mais complexifient la technique par concaténation (Dutoit, 1997). Dans le cadre de la synthèse par sélection d'unités, plusieurs inconvénients sont notables. L'élaboration des bases de données, par l'obtention des unités acoustiques et leur étiquetage, a nécessité un budget et un temps importants (Tabet & Boughazi, 2011). Les bases de données étant composées de plusieurs heures de parole enregistrée (Dutoit et al., 2002), un espace de stockage important est requis (Tabet & Boughazi, 2011). Il arrive aussi que certains segments soient mal étiquetés, aboutissant à une synthèse de la parole de mauvaise qualité (Tabet & Boughazi, 2011). Le signal de la synthèse par concaténation reste donc intelligible mais présente des lacunes en termes d'aspect naturel de la parole.

3.2. Synthèse articulatoire

Les méthodes de synthétisation de la parole présentées jusqu'alors comportent une limite importante dans le cadre de ce mémoire. Elles ne reposent sur aucune description du processus de production de la parole. Peu de contrôle est offert quant à certains paramètres importants pour une étude perceptive sur la parole, tels que des paramètres acoustiques (modélisation utilisée, fréquence fondamentale, etc.), géométriques (genre du locuteur, type de phonème, etc.) et articulatoires (vitesse d'articulation, etc.). L'utilisation de la synthèse articulatoire semble intéressante pour dépasser ces difficultés.

Présentée comme une méthode reflétant la réalité de la production de parole (Birkholz et al., 2006), la synthèse articulatoire consiste en une représentation modélisée de l'appareil phonatoire humain. Il s'agit d'un modèle géométrique et paramétré des articulateurs (Birkholz et al., 2006). La figure 7 expose le modèle de production élaboré par Kröger (1992), qui dépeint le processus dans lequel s'inscrit la synthèse articulatoire.

Pour générer un signal de parole, une conversion a lieu depuis une suite de symboles d'unités phonémiques (*« symbol entrance »*) vers un signal de parole acoustique (*« feature conversion »*). Chaque symbole est transformé en une caractéristique phonologique segmentaire distinctive (voyelle, semi-voyelle ou consonne). L'assemblage de ces caractéristiques forme la cible (*« target grid »*). Cette cible définit le moment où l'articulateur atteint l'objectif spatial visé. Par la suite, un modèle dynamique (*« dynamic model »*) est mis en jeu afin de contrôler la trajectoire des paramètres articulatoires (protrusion labiale, position linguale, ouverture vélaire, etc.) et phonatoires (ouverture glottique, tension cordale, pression pulmonaire) (*« control parameter trajectories »*). Ces paramètres articulatoires contrôlés sont ensuite transmis au modèle articulatoire (*« articulatory model »*). Celui-ci les traduit en diverses formes du tractus vocal (*« midsagittal shapes of the vocal tract »*). Une nouvelle conversion est alors réalisée. Elle permet de transposer les caractéristiques géométriques de l'appareil phonatoire modélisé en caractéristiques acoustiques (*« geometric – acoustic transformation »*). Enfin, une production du signal de parole acoustique (« *speech signal* ») est générée par le modèle physique (Kröger, 1992).

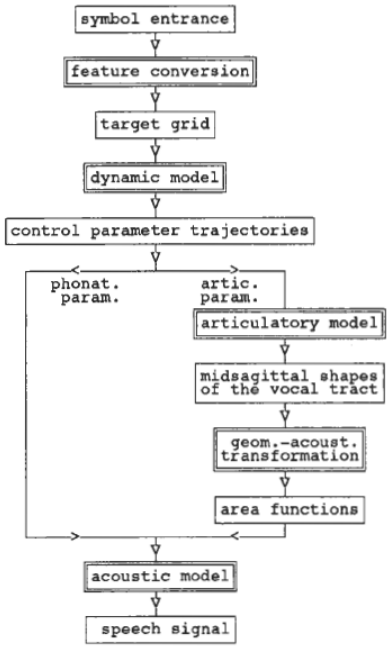

Figure 7 : Modèle de production : niveaux (rectangles simples) et modules (rectangles doubles) (Kröger, 1992)

Le modèle physique simule les processus physiques de la production de la parole. Concernant le modèle acoustique, différentes approches d'une complexité variable existent pour produire un signal de parole. Les **approches unidimensionnelles (1D)** peuvent être nommées. Largement explorées pour leur simplicité, elles considèrent que le champ acoustique résulte de la propagation d'ondes planes (Sondhi & Schroeter, 1987). Une onde plane est une unique variation de pression acoustique uniforme sur une section transverse du tractus vocal (R. Blandin, communication personnelle, 22 juin 2020). L'amplitude des ondes planes est indépendante de la courbure et de la forme de la section transverse du tractus vocal. Elle dépend seulement de l'aire de cette section transverse. Le tractus vocal peut donc être uniquement défini par sa variation d'aire transverse, ce qui constitue une représentation unidimensionnelle (Blandin et al., 2015 ; R. Blandin, communication personnelle, 6 août 2021). Cependant, l'approximation de la géométrie tridimensionnelle du tractus vocal par les modèles 1D reste limitée. Seules les fréquences inférieures à 5 kHz peuvent être décrites par des ondes planes uniquement. La validité de ces modèles se limite aux basses fréquences (Arnela et al., 2019 ; Freixes et al., 2018 ; Freixes et al., 2019). Or, le champ acoustique dans le tractus vocal se caractérise par des ondes planes et des résonances transverses, en hautes fréquences au-delà d'environ 5 kHz (Arnela et al., 2019 ; Blandin et al., 2015). Contrairement aux ondes planes, les résonances transverses correspondent à plusieurs variations de pression acoustique sur les sections transverses du tractus vocal (R. Blandin, communication personnelle, 22 juin 2020). Les **modèles tridimensionnels (3D)** dépassent la limite des approches 1D en représentant un champ acoustique tridimensionnel. En intégrant les résonances transverses des hautes fréquences en plus des ondes planes, ils offrent une description plus complète des effets de la géométrie tridimensionnelle du tractus vocal sur ses propriétés acoustiques (Arnela et al., 2019).

Les articulateurs sont le principal paramètre contrôlé dans la synthèse articulatoire. Ils regroupent l'ouverture labiale, la protrusion labiale, la position et la hauteur de l'apex lingual, la position et la hauteur de la langue et l'ouverture vélaire (Kröger, 1992). La notion d'« articulation » réfère à la mobilisation de toutes les structures impliquées dans la production de parole en termes de positions et de trajectoires (Scully, 1990). A travers la modulation de la position des articulateurs, les chercheurs visent à définir plus précisément le lien entre les changements anatomiques et les signaux acoustiques produits. L'objectif est également de mieux comprendre les modifications anatomiques du tractus vocal lors de la production de la parole (Huang, 2001). Il s'agit de créer un signal de parole à travers une simulation numérique du passage du flux d'air dans le tractus vocal. Ce type de synthèse cherche à intégrer des

dimensions neuromusculaires, articulatoires, aérodynamiques et acoustiques (Huang, 2001 ; Scully, 1990). Cette richesse dimensionnelle en fait potentiellement l'outil le plus efficient dans la génération de signaux de parole.

Employer les articulateurs comme paramètre principal permet de contrer les difficultés coarticulatoires et de concaténation auxquelles les autres méthodes de synthèse doivent faire face (Huang, 2001). Le lien avec le modèle physique aboutit à un signal plus réaliste. La synthèse articulatoire permet de personnaliser les signaux de parole par le contrôle de l'anatomie des articulateurs. Les attributs propres à un individu peuvent être modélisés. A titre illustratif, il est possible de modifier la longueur du tractus vocal (Huang, 2001). Le signal de tout type de locuteur peut être synthétisé, tel que celui d'un homme, d'une femme ou d'un enfant. Cette méthode permet aussi de synthétiser des voix aiguës ou graves et des voix dont la qualité est altérée dans le cadre d'une pathologie (Shadle & Damper, 2001). La synthèse articulatoire a l'avantage de ne pas être influencée par des effets liés au locuteur susceptibles d'influencer la production de parole (Delvaux & Pillot-Loiseau, 2019), ce qui fait d'elle un outil particulièrement adapté pour les expériences d'analyse perceptive (Birkholz et al., 2017).

La synthèse articulatoire n'échappe toutefois pas à certaines difficultés. En premier lieu, les connaissances relatives à la mobilité des articulateurs et leurs conséquences sont faibles. Généralement obtenues à l'aide de l'imagerie par résonance magnétique (IRM), elles ne permettent pas de définir les composants articulatoires les plus importants de la production de la parole à cause de leur grand degré de liberté (Freixes et al., 2019 ; Huang, 2001 ; Tabet & Boughazi, 2011). De nombreux éléments du tractus vocal doivent être modélisés et il n'est pas aisé de trouver un équilibre entre la facilité de contrôle et la précision lors de la construction du modèle articulatoire (Tabet & Boughazi, 2011). La contrôlabilité infinie des paramètres en fait sa force mais aussi sa faiblesse car le contrôle est plus complexe.

3.3. Réalisme de la synthèse articulatoire

Les méthodes de synthèse par formants et par concaténation ne paraissent pas entièrement satisfaire les exigences d'une synthèse de la parole idéale. D'après Arnela et al. (2019), la synthèse articulatoire semble posséder un grand potentiel. Elle est marquée par une plus grande flexibilité et une capacité à générer des signaux de parole avec des caractéristiques propres (Birkholz, 2013). Elle offre de nombreuses perspectives pour générer de la parole synthétique de qualité du point de vue de l'aspect naturel et de l'expressivité, grâce à son contrôle précis et flexible.

3.3.1. Aspect naturel du signal de parole synthétisé

En dépit des importantes avancées au cours des dix dernières années, Kuligowska et al. (2018) soulignent la nécessité de poursuivre le développement des systèmes de synthèse de parole à propos de l'aspect naturel des signaux de parole. Une différence perceptive persiste entre les signaux artificiels et ceux produits par un locuteur réel, malgré le fait que la synthèse de parole offre un signal naturel d'un point de vue acoustique (Kuligowska et al., 2018).

Produire un signal de parole synthétique proche d'un signal de parole humain est une tâche complexe. Plusieurs modèles doivent être impliqués, tels que les modèles physiologiques, aérodynamiques et acoustiques (Birkholz, 2013). Bien que la synthèse articulatoire ait connu d'importants progrès ces dernières années, la qualité du signal de parole n'est pas à la hauteur de celle du signal généré par la synthèse par sélection d'unités (Birkholz et al., 2017 ; Birkholz & Drechsel, 2021). Cependant, la synthèse articulatoire semble prometteuse pour obtenir un signal de parole artificielle proche d'un signal de parole naturel par l'emploi de modélisations et de simulations détaillées (Birkholz & Drechsel, 2021).

A l'aide d'un modèle 1D, Birkholz et Drechsel (2021) ont investigué l'influence de trois caractéristiques acoustiques du tractus vocal sur l'aspect naturel du signal de parole connectée : les sinus piriformes[3], le couplage acoustique transvélaire entre les cavités nasale et buccale et le rayonnement du son dans la peau du cou. Une première tâche de discrimination a été administrée, pour laquelle 22 participants (15 hommes et 7 femmes) ont été recrutés. Les stimuli étaient 10 mots allemands. Chacun était synthétisé en 8 combinaisons différentes, selon l'inclusion ou l'exclusion des caractéristiques acoustiques. L'objectif était de déterminer si les auditeurs étaient capables de discriminer deux combinaisons différentes d'un même mot. Les résultats ont révélé un taux de réponses correctes supérieur à un effet de hasard, indiquant l'habileté des participants à correctement discriminer des stimuli comportant ou non les caractéristiques acoustiques testées. Une seconde tâche perceptive a ensuite été proposée, qui consistait à évaluer l'aspect naturel de paires de mots. Les auditeurs devaient choisir le mot qui paraissait le plus naturel. L'hypothèse postulait que l'aspect naturel du signal de parole serait amélioré grâce à l'augmentation du réalisme du modèle par l'inclusion des trois caractéristiques susmentionnées. Les résultats ont infirmé cette hypothèse. Les trois caractéristiques acoustiques tendaient à diminuer l'aspect naturel des mots. Birkholz et Drechsel (2021) ont conclu que l'inclusion de ces caractéristiques avait accentué l'énergie en basse fréquence plutôt que celle

[3] Les sinus piriformes correspondent à deux cavités situées de part et d'autre du vestibule laryngée (Birkholz & Drechsel, 2021)

en haute fréquence. Le stimulus semblait plus « étouffé » d'un point de vue perceptif et donc, moins naturel.

La recherche de Birkholz et Drechsel (2021) met à jour le potentiel rôle des hautes fréquences pour produire un signal de parole plus naturel. L'usage de modélisations 3D, qui rendent compte de l'énergie acoustique dans les hautes fréquences, est peu répandu en raison de l'équipement informatique actuel qui rend leur application pratique difficile (Birkholz & Drechsel, 2021). Certains auteurs ont toutefois commencé à explorer la génération de sons synthétisés avec des modèles 3D (Arnela et al., 2019). Ils ont cherché à comparer la qualité de voyelles ([a], [i], [u]) et de diphtongues ([ɑi], [ɑu]) générées par des modèles 1D et des modèles 3D avec la synthèse articulatoire. Un résultat similaire entre les deux modèles a été produit dans les bandes fréquentielles basses. Cependant, l'étendue de la bande fréquentielle similaire dépend de la voyelle générée. Celle-ci était respectivement située à 3 kHz, 5 kHz et 6 kHz pour les voyelles [ɑ], [i] et [u]. Par ailleurs, une plus grande sensibilité aux changements de forme du tractus vocal a été démontrée pour le modèle 1D, qui ne parvient pas à estimer les fréquences des formants. D'après les auteurs, cela s'expliquerait par le fait que le modèle 1D s'appuie sur l'aire de chaque section transverse du tractus vocal. Contrairement à l'approche 3D qui corrige ce phénomène naturellement, l'approche 1D nécessite une correction manuelle (Arnela et al., 2019).

A ce jour, aucune étude ne semble avoir évalué les différences perceptives générées par l'usage des modèles 1D et 3D auprès d'auditeurs réels. En effet, les comparaisons menées étaient davantage centrées sur les techniques de génération des signaux de parole. De plus, seuls les phonèmes [ɑ], [i] et [u] ont fait l'objet d'une comparaison. Cette comparaison des signaux de parole générés par les modèles 1D et 3D ne représente qu'une petite partie des éléments phonétiques de la parole.

3.3.2. Expressivité du signal de parole synthétisé

La prosodie est une des principales limites rencontrées lors de la synthétisation de la parole. Elle réfère à la façon dont la phrase est prononcée du point de vue mélodique, rythmique, émotionnel et de l'accentuation (Kuligowska et al., 2018). Elle peut être porteuse d'informations paralinguistiques, relatives à l'état émotionnel du locuteur, ses caractéristiques propres (genre, âge, etc.), son style d'élocution et d'éventuels troubles vocaux ou de la parole (Birkholz et al., 2017). La qualité prosodique est souvent altérée dans la synthèse de parole en

raison des nombreux facteurs qui l'influencent, à savoir le sens de la phrase, les émotions et les caractéristiques du locuteur (Wagner, 2008, cité par Kuligowska et al., 2018). Tandis que les paramètres « primaires » de la prosodie (hauteur tonale, durée et intensité) sont facilement contrôlables dans la synthèse par concaténation, la manipulation des paramètres « secondaires » (type de phonation, longueur du tractus vocal, précision articulatoire et nasalité) est plus complexe avec ce type de synthèse (Kuligowska et al., 2018). Des modifications articulatoires simples peuvent en effet engendrer d'importantes difficultés acoustiques, qui peuvent être difficiles à gérer avec la synthèse par concaténation. Celle-ci voit également sa base de données limitée par rapport à l'ensemble des variations prosodiques possibles avec une voix humaine (Kuligowska et al., 2018). La synthèse par formants pourrait permettre de manipuler les paramètres « secondaires » de la prosodie par un contrôle temporal et spectral des caractéristiques de la source vocale et du tractus vocal. Kuligowska et al. (2018) indiquent cependant que la synthèse articulatoire est davantage appropriée pour aborder ce type de manipulation. Son étude s'est organisée autour de 9 mots allemands, produits par un locuteur natif, resynthétisés selon 7 variations différentes (tractus vocal long ou court, effort articulatoire faible ou fort, articulation légèrement ou très centralisée et nasalisation). Un total de 16 juges devait évaluer l'aspect naturel des mots et identifier le type de manipulation effectuée. Bien que les stimuli manipulés aient été considérés comme moins naturels que les stimuli non manipulés, les juges ont correctement identifié les paramètres contrôlés, excepté pour la précision articulatoire. A travers son étude, Kuligowska et al. (2018) ont montré qu'il était possible d'implémenter certains paramètres prosodiques « secondaires » dans des signaux de parole avec la synthèse articulatoire.

Birkholz et al. (2015) ont investigué l'influence du type de phonation sur la perception de 7 émotions différentes dans la parole (colère, ennui, dégoût, peur, joie, tristesse et neutralité). Ils rapportent le type de phonation comme l'un des paramètres les moins facilement manipulables dans la synthèse de parole. Trois courtes phrases en allemand, chacune produite avec les 7 émotions par 5 hommes et 5 femmes, ont été resynthétisées avec plusieurs types de phonation. Ceux-ci correspondaient à un type de phonation identique à celui de l'enregistrement réel, à une voix soufflée, modale ou pressée. La tâche a été proposée à 35 juges de 25.1 ans en moyenne (10 hommes et 25 femmes). Elle consistait à identifier l'émotion perçue dans la phrase à travers un choix forcé parmi les 7 émotions et à évaluer la force de cette émotion (de légère à forte). Les résultats ont montré la faculté des juges à correctement identifier les différentes émotions produites, sauf pour la peur. Il semblerait que certains types de phonation soient préférés pour des émotions particulières, à savoir une voix soufflée à modale pour la tristesse, une voix

modale pour la neutralité, le dégoût et la joie, une voix modale à pressée pour l'ennui, une voix pressée pour la colère et une voix soufflée pour la peur. Les résultats illustrent la possibilité de synthétiser de façon satisfaisante la majorité des émotions à l'aide de la synthèse articulatoire, par la modélisation d'un type de phonation particulier (Birkholz et al., 2015).

En résumé, la synthèse articulatoire a le potentiel de générer un signal de parole de haute qualité grâce à sa grande latitude de réglage de paramètres. La représentation modélisée de l'appareil phonatoire humain sur laquelle elle s'appuie permet d'étudier précisément le lien entre les changements anatomiques du tractus vocal et les signaux de parole produits. La précision et la flexibilité de son contrôle sont d'importants atouts pour générer un signal de parole plus naturel perceptivement et plus expressif par la manipulation d'aspects prosodiques. En particulier, l'emploi de modélisations 3D, qui tiennent compte de l'énergie acoustique dans les hautes fréquences de la parole, semble être prometteur pour contribuer à l'aspect naturel des signaux de parole artificiels.

III. OBJECTIFS ET HYPOTHESES

1. Présentation de l'objet d'étude

Ce mémoire s'inscrit dans un projet de développement d'une synthèse articulatoire à large bande. Il résulte d'une collaboration entre Angélique Remacle, chercheuse dans l'Unité Logopédique de la Voix de l'ULiège, et Rémi Blandin, chercheur dans l'Institut d'acoustique et de la parole de l'Université technique de Dresde. Ce projet vise à mieux comprendre et définir le lien entre la perception de la parole et les aspects physiques et acoustiques liés à sa production dans l'entièreté du registre fréquentiel audible (0.02 à 20 kHz). Pour ce faire, deux types de modélisation physique des HF sont utilisés : la modélisation 1D et la modélisation 3D.
Notre étude s'organise autour de deux objectifs, qui font chacun l'objet d'une expérience. Le premier objectif vise à évaluer la sensibilité auditive de jeunes adultes à divers degrés de réalisme dans la synthèse des HF, à l'aide d'une tâche de discrimination auditive dans un paradigme de comparaison par paires. Le second objectif cherche à déterminer si une description correcte de la physique dans la synthèse des HF, telle que dans le modèle 3D, contribue à l'aspect naturel de la parole artificielle, à l'aide d'une tâche d'évaluation de l'aspect naturel des stimuli via une échelle d'évaluation numérique.

2. Hypothèses de ce mémoire

A partir des données recueillies dans la littérature scientifique, plusieurs hypothèses sont émises concernant la relation entre la perception de la parole et la modélisation physique de l'énergie dans les HF. Les hypothèses relatives à la première expérience, suivies de celles en lien avec la seconde expérience, sont présentées ci-dessous.

2.1. Hypothèses concernant la première expérience

Nous souhaitons évaluer la sensibilité auditive des juges à des modélisations différentes des HF dans la parole synthétique. Diverses hypothèses sont testées concernant le type de paire de modèles (1D-1D ; 3D-3D ; 1D-3D), le genre de la voix de synthèse et le type de phonème. La fiabilité intra-juge entre le test et le retest et la fiabilité inter-juges sont aussi évaluées.

2.1.1. Le type de paire de modèles

A l'aide de la synthèse articulatoire, des phonèmes ont été générés à partir de modèles acoustiques d'une complexité différente : les modèles 1D et les modèles 3D. Les modèles 1D

possèdent des HF pour lesquelles la modélisation est physiquement incorrecte au-delà d'environ 5 kHz (Freixes et al., 2019). Les modèles 3D s'appuient sur l'entièreté du spectre audible (0.02 – 20 kHz) et offrent une modélisation physique plus réaliste des HF (Arnela et al., 2019). Les phonèmes ayant une composition dans les HF différente, une première hypothèse a été énoncée : **des différences seront perçues entre les phonèmes générés par le modèle 1D et ceux générés par le modèle 3D**. La confirmation de cette première hypothèse illustrerait la capacité des juges à percevoir une différence entre les deux modèles utilisés. En partant du postulat que les juges parviennent à percevoir une différence entre les modèles acoustiques, nous pourrions supposer qu'ils sont également capables de déterminer quand les modèles sont identiques dans une tâche de comparaison par paires. Nous nous attendons donc à ce que **les juges soient aussi performants pour comparer des paires de phonèmes générées à l'aide d'un même modèle acoustique (1D ou 3D) que pour comparer des paires de phonèmes générées à l'aide de deux modèles acoustiques différents (1D et 3D).**

2.1.2. Le genre de la voix de synthèse

La quantité d'énergie en HF dans les voix féminines est supérieure à celle dans les voix masculines pour des bandes d'octave et de tiers d'octave à 8 kHz, 10.1 kHz, 12.7 kHz, 16 kHz et 20.2 kHz (Monson et al., 2012). En considérant la répartition différente de l'énergie dans les HF selon le genre du locuteur et la composition acoustique de nos stimuli (amputés des fréquences inférieures à 5 kHz), nous avons énoncé l'hypothèse suivante : **les juges auront de meilleures performances pour comparer des paires de phonèmes générées avec une voix féminine plutôt qu'avec une voix masculine.**

2.1.3. Le type de phonème

En comparaison aux voyelles, les consonnes sont plus riches dans la gamme fréquentielle au-delà de 5 kHz (Monson, 2011). Monson, Hunter et al. (2014) précisent que la richesse en HF est d'autant plus visible au sein des consonnes fricatives [s], [ʃ], [f] et [θ]. Si l'intensité en HF est plus élevée dans les consonnes fricatives, nous supposons que la distinction entre une consonne fricative 1D et une consonne fricative 3D devrait être facilitée par rapport à la distinction de voyelles 1D et 3D. Nous nous attendons donc à ce que **les juges soient plus performants pour comparer des paires de consonnes fricatives non voisées ([s], [ʃ], [f]) générées avec deux modèles acoustiques différents (1D et 3D) que des paires de voyelles ([a], [i] [u] et [ə]) également générées avec deux modèles acoustiques différents (1D et 3D)**.

2.1.4. La fiabilité intra-juge

Un test-retest est introduit au sein de l'expérience. En présentant deux fois chaque paire, la fiabilité intra-juge des participants peut être vérifiée. Elle correspond au degré d'accord du participant envers lui-même quant à la réponse donnée. Nous nous attendons à ce que **les réponses des juges soient similaires au test et au retest.** De plus, ce test-retest permet de nous assurer de l'absence d'effet d'apprentissage chez les juges sélectionnés, qui consisterait en un biais pour nos analyses ultérieures.

2.1.5. La fiabilité inter-juges

Une hypothèse relative à la fiabilité inter-juges a été émise, qui stipule que **les réponses des différents juges seront similaires au sein de la tâche expérimentale proposée**. Elle vise à évaluer le degré de consensus des juges entre eux. Wuyts et al. (1999) conseillent en effet de considérer la fiabilité inter-juges lorsqu'une étude perceptive est menée.

2.2. Hypothèses concernant la deuxième expérience

Comme l'indiquent Monson, Hunter et al., « if humans can detect level changes (or absence vs. presence) of HFE in speech and voice, then HFE may contain information relevant to the percepts of speech and voice »[4] (p.6, 2014). La seconde expérience a pour but d'approfondir les résultats de la première expérience. Elle cherche à mieux définir la contribution des HF dans l'aspect naturel de phonèmes. Plusieurs hypothèses sont testées à propos du réalisme du modèle acoustique, du genre de la voix de synthèse et du type de phonème. Les fiabilités intra-juge et inter-juges sont également considérées.

2.1.1. Le réalisme du modèle acoustique

Une modélisation physique plus réaliste des HF est offerte par les modèles 3D (Arnela et al, 2019). Par ailleurs, le potentiel rôle des HF pour produire un signal de parole plus naturel a récemment été mis en lumière (Birkholz & Drechsel, 2021 ; Freixes et al., 2018). Nous nous attendons donc à ce que **les phonèmes générés par le modèle 3D soient perçus comme plus naturels que ceux générés par le modèle 1D**.

[4] « Si les hommes peuvent détecter des changements de niveau (ou l'absence par rapport à la présence) de HFE dans la parole et la voix, alors les HFE peuvent contenir des informations pertinentes pour la perception de la parole et de la voix. » (traduction personnelle)

2.1.2. Le genre de la voix de synthèse

Au regard des études menées sur la quantité d'énergie en HF dans le signal de parole en fonction du genre du locuteur, nous pourrions supposer que les phonèmes générés avec des voix féminines seraient associés à un plus haut degré de naturel. Cependant, les résultats de Monson et al. (2012) sont nuancés. Plutôt qu'une énergie supérieure ou inférieure en HF, la distribution des HF est différente entre les voix féminines et masculines. Les voix masculines sont plus riches dans la partie basse des HF alors que les voix féminines sont plus riches dans la partie haute des HF. Le spectre entier des stimuli étant transmis, les différences d'amplitude dans certaines fréquences selon le genre ne devraient pas induire de différence quant à la façon de percevoir les stimuli. De ce fait, l'hypothèse suivante est émise : **les phonèmes générés avec une voix féminine et ceux générés avec une voix masculine devraient être perçus avec le même degré de naturel.**

2.1.3. Le type de phonème

Compte-tenu la distribution différente de l'énergie en HF entre les voyelles et les consonnes (Monson, 2011), nous posons l'hypothèse que **le score attribué pour le degré de naturel variera selon les phonèmes**. En particulier, nous supposons qu'une plus grande quantité d'énergie en HF est susceptible d'accentuer la différence perçue entre les modèles acoustiques 1D et 3D. Les voyelles étant moins riches en HF, nous pensons que le changement de modélisation des HF (1D VS 3D) sera moins remarqué. Nous nous attendons donc à ce que **les consonnes fricatives non voisées [s], [ʃ], [f], générées avec le modèle 3D, soient considérées comme plus naturelles par les juges par rapport aux voyelles [a], [i] [u] et [ə] générées avec le modèle 3D.**

2.1.4. La fiabilité intra-juge

Comme dans la première expérience, la fiabilité intra-juge est évaluée par la double présentation des phonèmes. Celle-ci aboutit à un test-retest, qui nous permet de répondre à notre hypothèse qui stipule que **les résultats des juges seront équivalents au test et au retest.**

2.1.5. La fiabilité inter-juges

Une dernière hypothèse, qui suppose que **les résultats entre les juges seront équivalents concernant le score de degré de naturel attribué aux phonèmes**, a été émise pour évaluer la fiabilité inter-juges.

IV. METHODOLOGIE

1. Sélection des juges

1.1. Soumission du projet au comité d'éthique

La réalisation de l'étude dans le cadre de ce mémoire a reçu un avis favorable de la part du comité d'éthique de la Faculté de Psychologie, Logopédie et des Sciences de l'Education (FPLSE) en octobre 2020.

1.2. Modalités de recrutement

La procédure de recrutement des juges a eu lieu par l'emploi des réseaux sociaux, notamment par une annonce sur la plateforme Facebook. Le bouche à oreille a permis de diffuser la recherche à une plus large cohorte. Un e-mail de la part de Dominique Morsomme (responsable de l'Unité Logopédique de la Voix de l'ULiège) et d'Angélique Remacle auprès de leurs étudiants a également permis de recruter des juges supplémentaires. L'ensemble des outils utilisés conservaient une neutralité afin de ne pas entraver la liberté des juges d'accepter ou non de participer à l'étude.

1.3. Critères d'inclusion et d'exclusion des juges

Plusieurs critères d'inclusion et d'exclusion ont été définis afin de préciser le profil des juges recherchés.

Concernant les **critères d'inclusion**, les juges étaient des étudiant.e.s de 1ère et/ou de 2ème année de master en logopédie ou de master en psychologie à l'ULiège. En restreignant la sélection des juges à des étudiants en logopédie et en psychologie, nous cherchions à limiter l'influence des prototypes vocaux internes afin de conserver une homogénéité parmi les participants dans le cadre de notre étude perceptive (Papcun et al. 1989). Ces références prototypiques sont effectivement susceptibles d'influencer les stratégies perceptives employées (Kreiman & Gerratt, 1996). De telles limitations quant au profil des juges sont habituelles dans les études perceptives (Remacle et al., 2014). Chaque juge était âgé entre 20 et 29 ans car la sensibilité auditive aux HF tend à décliner avec l'âge. On observe une plus grande sensibilité auditive aux HF lorsque l'on est un jeune adulte par rapport à des personnes plus âgées (Monson, Hunter et al., 2014 ; Rodríguez Valiente et al., 2014 ; Zadeh et al., 2019). Le français était leur langue maternelle. Terbeek (1977, cité par Kreiman et al., 1990) a montré l'existence de différences entre les espaces perceptifs de participants dont la langue maternelle différait. Celle-ci

exercerait une influence sur les stratégies perceptives, de la même manière que le vécu de l'individu (Kreiman et al., 1990).

A propos des **critères d'exclusion**, les juges ayant rapporté des antécédents auditifs et/ou la présence de troubles auditifs étaient exclus de l'étude. Il semblerait que les enfants ayant connu des épisodes d'otites chroniques aient une audition significativement moins bonne dans les HF (Margolis et al., 2000). En revanche, si le juge présentait une audition normale au moment du testing (vérifiée par une audiométrie tonale (voir le point 2.2.), il n'était pas exclu de l'étude, malgré la mention d'antécédents auditifs. Les juges n'avaient pas de connaissances particulières dans le domaine de la synthèse de la parole pour éviter tout biais. Une expertise dans ce domaine peut aboutir à une adaptation de la stratégie perceptive employée (Bauer et al., 2010). Autrement dit, l'auditeur expert présente une performance significativement meilleure qu'un auditeur naïf.

Par ailleurs, nous avons tenu compte d'une potentielle expertise musicale chez nos juges car elle est susceptible d'améliorer la vitesse et la précision perceptive lors de modifications de hauteur tonale (Tervaniemi et al., 2005).

L'ensemble de ces variables est contrôlé par le biais d'un questionnaire anamnestique (voir l'annexe 2) que les juges ont complété en la présence de l'expérimentatrice avant de débuter le testing. Ainsi, tout questionnement du juge pouvait recevoir une réponse. Nous pouvions aussi vérifier la complétion correcte du questionnaire, qui a permis d'assurer l'homogénéité des juges et de pallier le problème de la voix prototypique divergente entre les individus. Une fiche de consentement éclairé a été signée par les juges avant la complétion du questionnaire anamnestique.

1.4. Description de l'échantillon final

Sur base d'études similaires d'analyse perceptive de la parole à l'aide de comparaison par paires (Baird et al., 2018 ; Kacha et al., 2005 ; Remacle et al., 2014), nous avions prévu de recruter entre 30 et 60 participants. Un pré-test sur 5 juges effectué en octobre 2020 a permis de préciser le nombre minimum de 30 de juges à recruter sur base d'un test statistique d'analyse de puissance. Pour ce faire, le logiciel G*power a été utilisé, avec une puissance de 80%.

Au total, 31 juges ont été recrutés dans le cadre de ce mémoire, dont 10 en master 1 (32.26%), 2 entre le master 1 et le master 2 (6.45%) et 19 en master 2 (61.29%). Ils étaient âgés de 21 à 28 ans, avec une moyenne d'âge de 22.9 ans et un écart-type de 1.51 an. La cohorte était majoritairement composée de femmes (87.10%) et d'étudiants en logopédie (70.97%).

Au sein de cet échantillon, 7 juges (22.58%) ont rapporté des antécédents auditifs, majoritairement caractérisés par des otites durant l'enfance. Parmi eux, deux juges ont signalé un épisode d'otite aux alentours de 20 ans. Un juge a précisé avoir des bouchons aux oreilles qui n'étaient toutefois pas présents au moment du testing et 2 juges ont dit avoir eu des acouphènes vers 18 ans de façon temporaire. Concernant l'expertise musicale, 17 juges (54.84%) ont indiqué pratiquer ou avoir pratiqué la musique et/ou le chant dans une académie de musique, par des cours particuliers ou en autodidacte. Parmi eux, 8 ont suivi des cours de solfège, 13 ont appris à jouer d'un instrument, 4 ont appris le chant lyrique, 6 le chant de variété et 6 le chant en chorale. Au moment de l'étude, 7 des 17 juges avaient arrêté leur pratique musicale. A propos des connaissances dans le domaine de la synthèse de la parole, 2 juges (6.45%) ont rapporté posséder des connaissances théoriques brèves et 19 (61.29%) ont indiqué en avoir déjà entendu parler. Les caractéristiques de la cohorte sont reprises dans le tableau 1.

	Total (n)	**Pourcentage calculé sur les 31 juges (%)**
Genre		
Homme	4	12.90
Femme	27	87.10
Année d'études		
Master 1	10	32.26
Master 1/2	2	6.45
Master 2	19	61.29
Filière des études		
Logopédie	22	70.97
Psychologie	9	29.03
Antécédents auditifs		
Non	24	77.42
Oui	7	22.58
Pratique musicale		
Non	14	45.16
Oui :	17	54.84
– Actuelle	10	32.26
– Passée	7	22.58
Connaissances dans le domaine de la synthèse de la parole		
Théoriques	2	6.45
Pratiques	0	0
Déjà entendu parler	19	61.29
	Moyenne	**Ecart-type**
Age	22.90	1.51

Tableau 1 : Caractéristiques de l'échantillon recruté

2. Description du matériel

2.1. Environnement de l'étude

Les testings ont été effectués dans la cabine audiométrique du CEDIA, qui est un bureau d'études et de recherche spécialisé en acoustique et vibrations (http://www.cedia.ulg.ac.be). Il est situé dans le quartier polytechnique du campus de Liège Sart-Tilman, à l'ULiège. La cabine était meublée d'un bureau, d'une chaise, d'un ordinateur ASUS X540Y et d'un haut-parleur XM6.D sn 07-3301 de la marque FAR by ATD. Des néons permettaient d'assurer l'éclairage de la pièce. Ceux-ci étaient néanmoins bruyants. Nous avons donc installé une lampe sur pied dans un coin de la cabine pour remplacer l'éclairage des néons et réduire le niveau du bruit de fond (voir figure 8). Un chauffage fixe était présent dans la cabine. Bien qu'il ait été éteint pendant chaque testing, un bruit continu de souffle était parfois émis, sans qu'il puisse être atténué. Le juge était invité à s'asseoir sur la chaise, face à l'ordinateur, pour réaliser les deux tâches perceptives.

Figure 8 : Système expérimental dans la cabine audiométrique du CEDIA

Préalablement aux pré-tests et aux testings, une mesure du bruit de fond du local a été réalisée pendant une minute à l'aide du sonomètre 01dB type Fusion sn 10601, calibré avec un calibrateur 01dB CAL21 sn 35293309. Le bruit de fond était de 22.8 dB SPL lorsque l'ordinateur était allumé. Des soucis informatiques rencontrés pendant la période de testing nous ont néanmoins amenés à changer d'ordinateur au profit d'un nouveau de la marque HP Pavilion 15-eh0076nb. Sur les 31 participants, 23 ont été testés avec le nouvel ordinateur. Celui-ci étant récent, la ventilation était très silencieuse. De nouvelles mesures de bruit de fond ont été réalisées qui étaient de 18.4 dB SPL. Afin de conserver des conditions environnementales identiques à celles pour les premiers juges testés, nous avons décidé d'utiliser les deux ordinateurs. Ainsi, l'ancien ordinateur était allumé sur le bureau pour avoir le même niveau de

bruit de fond que pour les premiers testings. Le nouvel ordinateur, également posé sur le bureau, était utilisé pour réaliser les tâches perceptives. La figure 9 représente le système expérimental adapté.

Figure 9 : Système expérimental adapté dans la cabine audiométrique du CEDIA

Les stimuli étaient délivrés à 70 dB SPL à travers le haut-parleur. Il était placé en face de l'individu, à 1 mètre de distance (comme dans les études de Monson et al., 2011 ; Monson, Hunter et al., 2012 ; Monson & Caravello, 2019). Son niveau d'amplification était de -6.5 dB pour ajuster le niveau de pression sonore[5] des stimuli à 70 dB SPL. Monson, Lotto et al. (2014) ont indiqué que la propagation des HF se marque, en autres, par des effets de diffusion et de direction. A mesure que la fréquence augmente, le rayonnement de l'énergie acoustique devient de plus en plus directionnel. Par conséquent, une attention particulière a été portée à la position du haut-parleur dans le cadre de notre étude. Il était demandé au participant de bouger le moins possible depuis sa position assise et de ne pas se pencher vers le haut-parleur, pour éviter toute déviation par rapport à celui-ci (Monson, Lotto et al., 2014).

2.2. Audiométrie tonale

Une audiométrie tonale a été réalisée chez les juges pour évaluer leurs seuils auditifs. Sa durée était d'environ 20 minutes. Elle a été administrée avec l'audiomètre Madsen Itera II, qui permet l'évaluation des seuils auditifs de fréquences jusqu'à 20 kHz (voir figure 10). N'ayant toutefois été étalonné que jusqu'à 8 kHz en 2019 (voir annexe 3), seules les fréquences 0.5, 1, 2, 4 et 8 kHz ont été testées avec cet audiomètre. D'après la classification audiométrique des déficiences auditives du Bureau International d'Audiophonologie (BIAP) (n.d.), l'audition normale se définit par une perte auditive qui n'excède pas 20 dB. Le juge était donc exclu de

[5] Le niveau de pression sonore correspond à l'intensité ou au volume auquel les stimuli sont délivrés. Il est exprimé en décibels (dB).

l'étude si une des fréquences de 0.5 à 8 kHz, pour au moins une des deux oreilles, présentait un seuil auditif supérieur à 20 dB HL (Monson & Caravello, 2019).

Figure 10 : Audiomètre Madsen Itera II (à gauche) et casque (à droite)

En vue de répondre à nos questions de recherche, les HF à 12.5 et 16 kHz ont été testées à l'aide d'un audiomètre développé au CEDIA par Xavier Kaiser, ingénieur de recherches à l'ULiège (voir figure 11). Le casque Sennheiser HDA 300 a été utilisé. Des difficultés techniques ont cependant altéré la fiabilité de la fréquence à 16 kHz. Les juges ne percevaient pas tous le même stimulus. Certains entendaient un son continu sans bruit parasite tandis que d'autres entendaient un bruit parasite au début et à la fin du stimulus, sans entendre le stimulus. La majorité des juges confrontés à ce problème technique ne répondaient d'ailleurs pas à la stimulation auditive, pensant que le bruit parasite était un grincement du casque. La fréquence de 16 kHz n'a donc pas été considérée pour l'inclusion des juges dans l'étude.

Figure 11 : Audiomètre développé par le CEDIA (à gauche et au centre) et casque Sennheiser HDA 300 (à droite)

L'audiométrie tonale a permis d'apprécier l'acuité auditive des juges, qui est liée à la détection de l'énergie des HF au sein de la parole (Monson & Caravello, 2019). En outre, il était précisé aux juges que l'audiométrie n'avait pas de visée diagnostique, c'est pourquoi leurs résultats ne leur étaient ni montrés, ni transmis. Ils devaient réaliser la totalité de l'expérience même si leurs seuils auditifs étaient mauvais. Aucun juge n'a toutefois présenté de mauvais seuils auditifs et été exclu de l'étude. Les résultats de l'audiométrie tonale de chaque juge sont présentés dans l'annexe 4.

2.3. Stimuli utilisés

Les stimuli ont été modélisés par Rémi Blandin à l'aide du logiciel VocalTractLab (http://vocaltractlab.de/) qui propose un synthétiseur articulatoire de la parole. Mis au point par Peter Birkholz, cet outil permet le contrôle d'un modèle géométrique 3D du tractus vocal humain à travers la manipulation de la forme et de la position de 23 paramètres vocaux (voir figure 12). Ces paramètres concernent la protrusion labiale, l'angle d'ouverture de la mâchoire, la position du voile du palais ou encore l'élévation linguale. Les dimensions articulatoire, acoustique et physique sont prises en compte. Grâce à une représentation intuitive et flexible des articulateurs, une meilleure compréhension des processus de production de la parole est possible (Birkholz et al., 2006). Le modèle acoustique 3D a été implémenté dans le logiciel utilisé. En outre, la bande passante des stimuli était large (de 0.02 à 20 kHz) afin de couvrir l'ensemble des fréquences audibles par l'être humain et de tester, en particulier, le rôle des HF.

Sept phonèmes synthétiques monotones d'une durée de 670 ms ont été utilisés pour les deux expériences. Il y avait 4 voyelles ([a], [i], [u], [ə]) et 3 consonnes fricatives non voisées ([f], [s], [ʃ]). Leur choix est justifié par leur variabilité, bien que leur nombre soit réduit. L'énergie acoustique contenue dans les HF diffère en effet selon qu'il s'agisse de voyelles ou de consonnes (Monson & Buss, 2019). La composition acoustique des phonèmes est différente entre la première et la seconde expérience. Les points 2.3.1. et 2.3.2. détaillent ces différences.

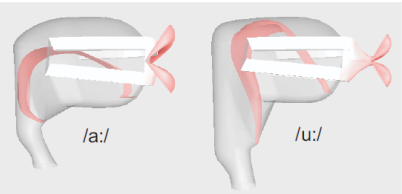

Figure 12 : Géométrie du tractus vocal pour les voyelles [a:] et [u:]
(Birkholz et al., 2006)

2.3.1. Stimuli utilisés pour la première expérience

La première expérience se fonde sur des comparaisons de paires de phonèmes, pour lesquels la synthétisation physique est identique ou différente. Chaque phonème a été modélisé avec le modèle 1D et le modèle 3D. Trois types de comparaison par paires ont été formés :

- la comparaison de phonèmes 1D-1D (par exemple, [a] du modèle 1D et [a] du modèle 1D)
- la comparaison de phonèmes 3D-3D (par exemple, [a] du modèle 3D et [a] du modèle 3D)
- la comparaison de phonèmes 1D-3D (par exemple, [a] du modèle 1D et [a] du modèle 3D)

Tous les stimuli ont été modélisés selon les genres masculin et féminin et ont été présentés deux fois, qui correspondent aux moments du test et du retest. Chaque type de comparaison était composé de 28 paires de stimuli (7 phonèmes * 2 genres * 2 moments). Au total, 84 paires de stimuli ont été présentées (7 phonèmes * 2 genres * 2 moments * 3 types de comparaison).
La question de l'implémentation d'un bruit de masque dans les basses fréquences des stimuli s'est posée et a été résolue à l'aide de pré-tests auprès de 6 juges externes (4 hommes et 2 femmes) à la cohorte recherchée, âgés de 21 à 36 ans. Le bruit de masque consiste à masquer un bruit plus faible par un bruit plus fort. Dans un premier temps, un bruit de masque dans les basses fréquences des phonèmes, avec une forme similaire à de la parole, a été ajouté, comme dans l'étude de Vitela et al. (2015). L'ajout d'un bruit de masque permettait d'assurer le fait que les juges évaluaient uniquement les HF et qu'ils ne s'appuyaient pas sur des otoémissions acoustiques provoquées pouvant être présentes dans les basses fréquences (Vitela et al., 2015). Cependant, les résultats des pré-tests et des questions posées aux juges ont révélé que leur attention était portée sur le bruit de masque et non sur les HF des phonèmes pendant la tâche perceptive. Pour rappel, plus la fréquence est élevée dans le spectre de la parole, plus l'intensité moyenne du spectre diminue (Monson et al., 2012). Les HF sont donc moins intenses que les basses fréquences dans la parole humaine. Le bruit de masque couvrant les basses fréquences, il tendait à avoir un plus haut niveau de pression sonore que les HF. Il a donc été retiré pour éviter qu'il soit un biais dans l'étude perceptive. Pour être sûr que les HF seules soient évaluées, les basses fréquences des phonèmes de 0 à 5.6 kHz ont aussi été retirées. Les juges n'entendaient ainsi que les HF des phonèmes durant la tâche perceptive.

2.3.2. Stimuli utilisés pour la seconde expérience

La deuxième expérience s'appuie sur les mêmes phonèmes que pour la première expérience ([a], [i], [u], [ə], [f], [s], [ʃ]) mais leur composition acoustique est différente. L'objectif de la seconde tâche expérimentale étant l'évaluation de l'aspect naturel des phonèmes, il s'agissait de présenter des phonèmes semblables à de la parole naturelle. Dès lors, conserver des phonèmes identiques à la première tâche expérimentale, sans basses fréquences, n'était pas possible. Nous ne sommes jamais confrontés, dans la vie quotidienne, à des signaux de parole amputés de leurs basses fréquences. Ainsi, les phonèmes de cette seconde expérience étaient composés de leur spectre acoustique complet, sans que leurs basses fréquences ne soient filtrées.
Comme pour la première expérience, chaque phonème a été modélisé à l'aide du modèle 1D et du modèle 3D et a été généré selon les genres masculin et féminin. L'ensemble des phonèmes

a été présenté deux fois, au moment du test et au moment du retest. Au total, 56 stimuli synthétiques ont été utilisés (7 phonèmes * 2 modèles * 2 genres * 2 moments).

3. Procédure expérimentale

Cette section aborde la manière dont s'est déroulée notre étude (voir annexe 5). Une seule rencontre avec chaque juge a eu lieu, d'une durée approximative d'1 heure et 15 minutes. Tous les juges ont été testés individuellement entre mars et mai 2021. La procédure expérimentale s'est organisée autour de deux tâches perceptives, qui sont détaillées dans les points 3.2. et 3.3. ci-dessous.

3.1. Déroulement général de l'étude

A leur arrivée, les juges étaient invités à compléter le consentement éclairé qui leur stipulait qu'ils étaient libres de mettre fin à l'expérience à tout moment, sans se justifier. L'addendum consentement procédure COVID était également signé, qui indiquait qu'ils seraient avertis si l'expérimentatrice était testée positive au Covid dans les jours suivants le testing. Le questionnaire anamnestique était ensuite complété. Puis, l'audiométrie tonale était administrée, suivie des deux tâches perceptives. Une pause de 5 minutes était systématiquement proposée aux juges entre l'audiométrie tonale et les tâches perceptives. Une cabine audiométrique est un environnement clos et sans fenêtre, auquel nous sommes peu habitués. Il était donc possible que les juges se sentent mal à l'aide ou éprouvent un inconfort dans la pièce. Ils pouvaient ainsi prendre l'air s'ils en ressentaient le besoin. La figure 13 expose les étapes de la procédure expérimentale générale d'un point de vue chronologique.

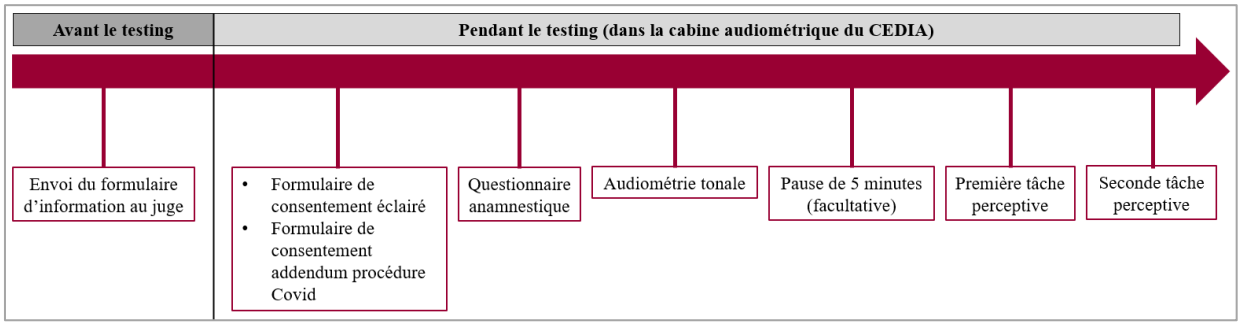

Figure 13 : Etapes de la procédure expérimentale selon un axe temporel

3.2. Première tâche expérimentale

La première tâche expérimentale consistait en une épreuve perceptive de comparaison par paires. Le juge était amené à écouter deux phonèmes puis à répondre à la question suivante : « les deux sons sont-ils identiques ou différents ? ». Afin de répondre à cette question, un ordinateur et une souris étaient à disposition du juge. Ce dernier observait sur l'écran une interface (voir figure 14) qui exposait deux boutons (son 1 et son 2), permettant d'écouter chacun des sons. Différentes paires de phonèmes étaient générées aléatoirement : des paires différentes (1D-3D) et des paires identiques (1D-1D et 3D-3D). L'ordre des paires différentes était également aléatoire, c'est-à-dire que la paire pouvait être présentée dans l'ordre 1D-3D ou 3D-1D. Les paires de phonèmes ont toutes été testées durant la même session. Le juge était autorisé à écouter chaque son autant de fois qu'il le souhaitait, jusqu'à ce qu'il parvienne à statuer sur la présence ou l'absence d'une différence perceptive. Puis, il était invité à cliquer sur le bouton « identique » ou « différent » situé juste en-dessous et à passer à la paire suivante. La possibilité d'écouter à plusieurs reprises les sons étant offerte, aucun retour en arrière n'était permis.

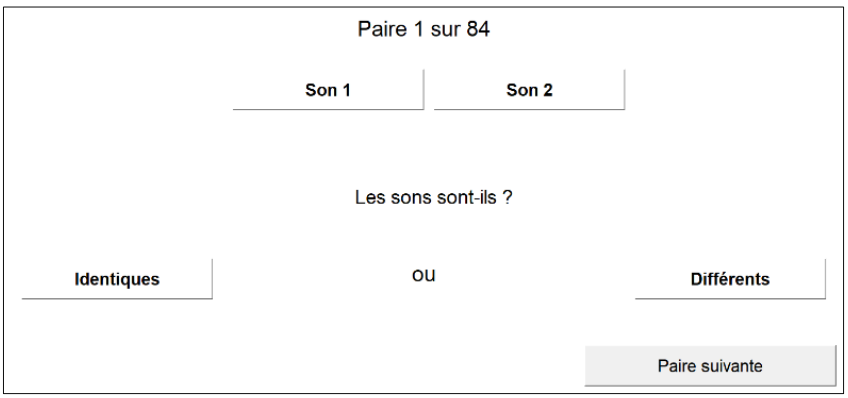

Figure 14 : Interface utilisée pour la première tâche expérimentale

Afin de préciser ce que nous entendions par les termes « identiques » et « différents », une phase d'entraînement a été introduite dans notre procédure expérimentale. Dans un premier temps, une paire de sons identiques et une paire de sons différents étaient jouées avant de débuter la tâche perceptive. Les paires jouées étaient les mêmes pour tous les juges pour qu'ils aient les mêmes références. Il s'agissait des comparaisons 1D-1D et 1D-3D avec le phonème [u] masculin. Dans un second temps, le juge était amené à réaliser la tâche perceptive avec une paire de sons 1D-1D avec une consonne proche du [ʃ] et une paire de sons 3D-1D avec le phonème [ɛ] proposées en tant qu'exemples. Les phonèmes de la phase d'entraînement étaient volontairement différents des 7 phonèmes inclus dans le testing pour éviter de perturber

l'expérience. L'objectif était que le juge se familiarise avec l'interface. La durée de la tâche était d'environ 20 minutes.

La consigne, donnée oralement aux juges, était la suivante : « *Vous allez écouter 2 sons et vous devrez indiquer si vous avez l'impression qu'ils sont identiques ou différents. Voici à quoi ressemble une paire de sons identiques (jouée par l'expérimentatrice) et voici à quoi ressemble une paire de sons différents (jouée par l'expérimentatrice). Il peut s'agir de tout type de différence. Pour écouter les sons, vous devez cliquer sur le bouton « son 1 » et le bouton « son 2 ». Pour passer à la paire suivante, vous devez cliquer sur « paire suivante ». Les sons pourront vous paraître particuliers car ce sont des sons que nous n'avons pas l'habitude d'entendre dans notre quotidien. Vous pouvez les écouter autant de fois que vous le souhaitez mais lorsque vous aurez cliqué sur « paire suivante », il ne sera plus possible de revenir en arrière. Nous allons d'abord faire un entraînement avec 2 paires de sons, puis vous aurez 84 paires de sons à juger. Si vous avez des questions à la fin de cet entraînement, n'hésitez pas à les poser. ».*

D'après plusieurs auteurs, le paradigme de comparaison par paires est une procédure adaptée pour l'évaluation de signaux complexes (Larrouy-Maestri, in press ; Kacha et al., 2005). L'emploi de jugements comparatifs tend à augmenter les fiabilités intra- et inter-juges pour l'évaluation de voyelles soutenues et de phrases, qu'il s'agisse de juges experts ou non (Kacha et al., 2005 ; Teston, 2004). Ce type de tâche offre un point de comparaison commun entre les participants, ce qui évite le recours à leurs standards internes (Remacle et al., 2014). Afin de tester la fiabilité intra-juge, chaque paire était présentée deux fois. Nous avons ainsi pu vérifier la concordance des réponses de chacun des participants à deux moments : le test et le retest. La fiabilité inter-juges était évaluée par l'administration d'une tâche identique, comportant les mêmes stimuli, à tous les juges (Remacle et al., 2014). L'ordre de présentation des paires de phonèmes était randomisé pour chaque juge afin de contrer un effet de comparaison du stimulus par rapport au précédent. Les données recueillies ont automatiquement été enregistrées dans un fichier Excel.

3.3. Seconde tâche expérimentale

La seconde tâche expérimentale consistait à écouter les 56 phonèmes synthétiques et à évaluer leur aspect naturel. Chaque phonème était proposé individuellement. Une interface comportant une échelle de Likert était présentée sur l'écran d'ordinateur (voir figure 15). Cette échelle était initialement graduée de 0 (« pas du tout naturel ») à 7 (« totalement naturel »). Le

juge devait cliquer avec la souris sur le chiffre qui lui semblait représenter au mieux le degré de naturel du phonème écouté. Des pré-tests sur 4 participants (3 hommes et 1 femme) âgés de 21 à 24 ans nous ont amenés à modifier la gradation de l'échelle de Likert. Ils trouvaient l'échelle trop large et rencontraient des difficultés à déterminer un chiffre représentatif de l'aspect naturel du phonème entendu. D'après Tullis et Albert (2013), ce constat n'est pas étonnant. Plus l'échelle comporte de chiffres, plus il est difficile d'y associer des termes descriptifs. Bien que l'échelle de Likert en 5 points soit généralement utilisée (Tullis & Albert, 2013), nous avons choisi de réduire la gradation à 4 points, de 0 (« pas du tout naturel ») à 3 (« totalement naturel »), afin d'annuler la possibilité d'opter pour une réponse neutre. Comme dans beaucoup d'autres études utilisant une échelle de Likert, seuls les échelons extrêmes (0 et 3) ont été étiquetés (Tullis & Albert, 2013). De nouveaux pré-tests auprès de 3 des 4 participants (3 hommes de 21 ans) ont soutenu ce choix puisqu'ils parvenaient plus facilement à statuer sur l'aspect naturel du phonème. A nouveau, les juges pouvaient écouter le phonème autant de fois que désiré, sans pouvoir retourner en arrière.

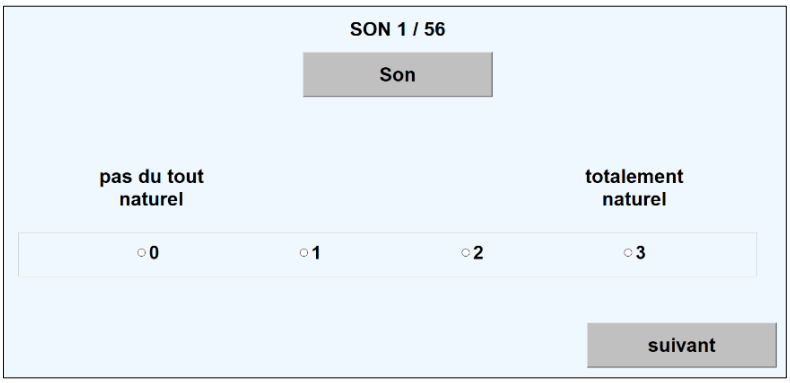

Figure 15 : Interface utilisée pour la seconde tâche expérimentale

L'ensemble des juges durant les pré-tests ont exprimé leur incompréhension et doute face à la signification du terme « aspect naturel ». Nous avons donc précisé dans la consigne que ce terme fait référence à une voix que nous pouvons entendre dans la vie de tous les jours. Une phase d'entraînement, composée d'une consonne 1D proche du [ʃ] et du phonème [ɛ] 3D, a été implémentée pour que le juge se familiarise avec l'interface. A nouveau, nous avons choisi des phonèmes différents des 7 phonèmes inclus dans le testing pour ne pas perturber l'expérience perceptive. La durée de cette seconde tâche était d'environ 10 minutes.

La consigne, donnée à l'oral, était la suivante : « *Vous allez écouter des voyelles et des consonnes et vous devrez évaluer leur aspect naturel sur une échelle de 0 à 3, 0 signifiant « pas*

du tout naturel » et 3 signifiant « totalement naturel ». Par aspect naturel, nous faisons référence à une voix que nous pouvons entendre dans la vie de tous les jours. Il n'y a pas de bonne ou de mauvaise réponse, c'est selon votre ressenti. Pour écouter le son, vous devez cliquer sur le bouton « son ». Pour passer au son suivant, vous devez cliquer sur « suivant ». Vous pouvez écouter le son autant de fois que vous le souhaitez mais lorsque vous aurez cliqué sur « suivant », il ne sera plus possible de revenir en arrière. Nous allons d'abord faire un entraînement avec 2 sons, puis vous aurez 56 sons à évaluer. Si vous avez des questions à la fin de cet entraînement, n'hésitez pas à les poser. ».

La fiabilité intra-juge a été vérifiée par un test-retest, à travers la double présentation de chaque phonème (Remacle et al., 2014). Les mêmes phonèmes ont été présentés à tous les juges. Leur ordre de présentation était aléatoire pour chaque juge. L'ensemble des données recueillies a automatiquement été enregistré dans un fichier Excel.

4. Analyses statistiques

Les analyses statistiques de ce mémoire ont été réalisées par Vincent Didone, logisticien de recherche à l'ULiège. Elles ont permis de tester les différentes hypothèses formulées précédemment à propos de la première et de la seconde tâche expérimentale.

4.1. Analyses statistiques réalisées pour la première tâche expérimentale

Les analyses statistiques de la première expérience perceptive ont été réalisées à l'aide du package lme4, sur les données du moment du test. Le modèle statistique utilisé est une **régression linéaire mixte généralisée de type binomiale**. Les réponses récoltées auprès des juges étaient initialement codées en 0 et 1. Le score de 1 signifiait que la paire de phonèmes était considérée comme identique et le score de 0 signifiait que la paire était considérée comme différente. Néanmoins, ce codage n'offrait pas d'indication quant au fait que la réponse était correcte ou non. Un recodage des données a donc eu lieu de façon à parler en termes de réponse correcte ou incorrecte plutôt qu'en termes de réponse identique ou différente. A titre d'exemple, si la paire 1D-3D était considérée comme différente, le score de 1 était associé à une réponse correcte. Si la paire 1D-3D était considérée comme identique, le score de 0 était associé à une réponse incorrecte. Par le recodage de données nominales en données linéaires, un score de discrimination correcte ou incorrecte a été obtenu. Le modèle de régression linéaire mixte généralisée permet également d'introduire un effet aléatoire lié aux juges. Il permet l'obtention de droites de régression différentes entre les juges et ainsi d'évaluer la fiabilité inter-juges.

Celle-ci a été interprétée à l'aide d'un **coefficient de corrélation intra-classe (ICC)** qui est une mesure largement utilisée pour évaluer le degré d'accord entre juges (Koo & Li, 2016).

Plusieurs variables dépendantes et indépendantes ont été implémentées dans le modèle en vue de répondre à nos hypothèses : le score de réussite (paire bien discriminée ou pas), l'effet aléatoire du juge, le moment (test ou retest), le type de paire de modèles (1D-1D, 3D-3D ou 1D-3D), le type de phonème ([a], [i], [u], [ə], [f], [s] ou [ʃ]) et le genre de la voix de synthèse (féminin ou masculin).

Ces variables réfèrent aux **effets simples** de l'étude. Un test de Wald, basé sur une distribution du X^2, a été réalisé pour mettre en évidence la significativité de certains effets simples testés. Afin d'approfondir la significativité des résultats, des contrastes linéaires ont été menés. Il s'agit de comparaisons multiples. Un contrôle de l'erreur de l'alpha de première espèce a été réalisé à l'aide de la méthode de Holm pour l'ensemble des comparaisons. En effet, plus le nombre de comparaisons réalisées est grand, plus il y a de chance de mettre en évidence un effet significatif qui n'est pas réel. Cet effet significatif correspond à l'erreur de première espèce. Par son contrôle, nous obtenons des valeurs p ajustées. Les scores des juges ayant été normalisés, les analyses se sont appuyées sur une distribution des scores z.

Plusieurs types de corrélation ont également été introduits, qui correspondent aux **effets d'interaction** de l'étude : type de phonème * type de paire de modèles ; type de paire de modèles * genre de la voix de synthèse ; genre de la voix de synthèse * type de phonème. Les deux derniers effets d'interaction cités font l'objet d'analyses statistiques complémentaires car nous n'avons pas formulé d'hypothèse à leur propos.

4.2. Analyses statistiques réalisées pour la seconde tâche expérimentale

Les analyses statistiques de la seconde expérience perceptive ont été réalisées à l'aide d'une **régression logistique cumulative**, sur les données du moment du test. Elle est **de type ordinal** car les données récoltées au cours de la seconde tâche expérimentale sont ordinales. Pour rappel, le mode de réponse des juges était une échelle de Likert en quatre points, de 0 « pas du tout naturel » à 3 « totalement naturel ». Le modèle de distribution ordinale a été ajusté avec l'approximation de Laplace.

Dans un premier temps, un **test du rapport de vraisemblance des modèles de liens cumulatifs** a été réalisé afin de tester les effets simples. Il se traduit par une comparaison de modèle statistique à modèle statistique à travers l'ajout des variables indépendantes une à une à partir d'un modèle complètement vide. Ce dernier ne comportait que la variable dépendante

(les scores d'évaluation de l'aspect naturel des phonèmes) et l'effet aléatoire des juges. Les comparaisons ont permis d'évaluer les effets significatifs sans prendre en considération les catégories à l'intérieur de ces effets. A titre d'exemple, l'effet du phonème était considéré pour l'ensemble des phonèmes confondus et pas pour un phonème particulier. La figure 16 représente la procédure statistique. A chaque niveau du modèle, une nouvelle variable est ajoutée. L'effet du moment a d'abord été testé, suivi de l'effet du réalisme du modèle acoustique, du type de phonème et du genre de la voix de synthèse. L'interaction entre le réalisme du modèle acoustique et le type de phonème et celle entre le réalisme du modèle acoustique et le genre de la voix de synthèse ont ensuite été intégrées. La significativité des résultats a été interprétée à l'aide d'une distribution du X^2.

```
clmm.6   Evaluation ~ 1 + (1 | Juge)
clmm.5   Evaluation ~ Moment + (1 | Juge)
clmm.4   Evaluation ~ Moment + Model + (1 | Juge)
clmm.3   Evaluation ~ Moment + Model + Phoneme + (1 | Juge)
clmm.2   Evaluation ~ Moment + Model + Phoneme + Genre + (1 | Juge)
clmm.1   Evaluation ~ Moment + Model + Phoneme + Genre + Model:Phoneme + (1 | Juge)
clmm.max Evaluation ~ Moment + Model + Phoneme + Genre + Model:Phoneme + Model:Genre
```

Figure 16 : Procédure statistique du test du rapport de vraisemblance des modèles de liens cumulatifs

Dans un second temps, des **contrastes linéaires** au niveau des variables ont été réalisés en vue d'investiguer plus en profondeur les effets significatifs. Elles consistent à comparer les niveaux de la variable entre eux pour situer la significativité de l'effet. La significativité des résultats a été interprétée à l'aide d'une distribution des scores z.

Plusieurs variables indépendantes ont été introduites dans le modèle, qui correspondent à des **effets simples** : le réalisme du modèle acoustique (1D ou 3D), le genre de la voix de synthèse (féminin ou masculin), le type de phonème ([a], [i], [u], [ə], [f], [s] ou [ʃ]) et le moment (test ou retest). La variable dépendante du score d'évaluation de l'aspect naturel du stimulus a aussi été implémentée. Des corrélations ont été introduites, qui réfèrent aux **effets d'interaction** : réalisme du modèle acoustique * type de phonème ; réalisme du modèle acoustique * genre de la voix de synthèse. Ce dernier effet d'interaction fait l'objet d'analyses statistiques complémentaires car nous n'avons pas émis d'hypothèse à son propos. Un effet aléatoire lié aux juges a été ajouté au sein du modèle statistique pour tenir compte de la fiabilité inter-juges. Celle-ci a été interprétée à l'aide d'un **ICC**.

V. RESULTATS

1. Résultats des hypothèses

Les résultats des hypothèses relatives à la première expérience de comparaison par paires sont présentés dans le point 2.1. Les effets simples testés sont d'abord abordés, suivis des effets d'interaction. Les résultats des hypothèses relatives à la deuxième expérience d'évaluation de l'aspect naturel des phonèmes sont exposés dans le point 2.2. Les résultats des effets simples sont présentés en premier lieu, suivis des résultats des effets d'interaction en second lieu.

2.1. Résultats des hypothèses de la première tâche expérimentale

2.1.1. Résultats des hypothèses des effets simples

La figure 17 expose le graphique général, qui regroupe l'ensemble des variables indépendantes introduites dans le modèle statistique. Le type de phonème, le genre de la voix de synthèse (avec 0 = féminin et 1 = masculin) et le type de paire de modèles (avec les paires 1D-1D en rouge, 1D-3D en gris et 3D-3D en orange) sont indiqués sur l'axe des abscisses. L'axe des ordonnées regroupe le moment (test ou retest) ainsi que le pourcentage de réussite de discrimination des paires de phonèmes. Ce graphique est abordé de manière plus détaillée dans les points 2.1.1.1., 2.1.1.2. et 2.1.1.3., à travers la considération des variables une à une.

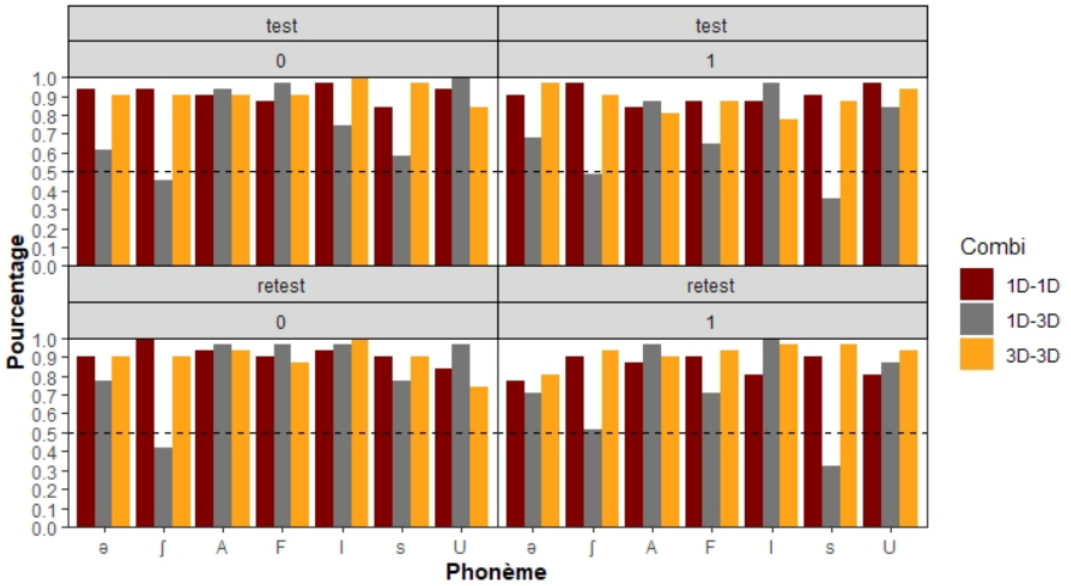

Figure 17 : Graphique comportant les résultats de l'ensemble des variables indépendantes testées

Le tableau 2 reprend les données statistiques de la régression linéaire mixte généralisée de façon chiffrée pour les effets simples. Les données surlignées en gras correspondent aux résultats reflétant une différence significative ($p \leq 0.05$). Elles sont détaillées dans les points qui suivent.

Variable indépendante	X^2	Degré de liberté (Df)	Probabilité (> X^2)
Type de paire de modèles	76.819	2	**< 0.000**
Genre de la voix de synthèse	10.490	1	**< 0.001**
Moment	1.449	1	> 0.1
Type de phonème	46.62	6	**< 0.000**

Tableau 2 : Résultats de la régression linéaire mixte généralisée pour les effets simples

2.1.1.1. Hypothèse concernant l'effet du type de paire de modèles

Tout d'abord, nous avons évalué l'influence du type de paire de modèles sur la capacité des juges à discriminer des paires de sons. Pour rappel, notre première hypothèse stipule que **des différences seront perçues entre les phonèmes générés par le modèle 1D et ceux générés par le modèle 3D**. La figure 18 représente le pourcentage de réussite des 31 juges à la tâche de discrimination de paires de phonèmes, pour chaque type de paire. Le pourcentage de réussite est indiqué sur l'axe des ordonnées, tandis que les types de paire de modèles sont présentées sur l'axe des abscisses. Le score de discrimination des paires 1D-1D est représenté en rouge, celui des paires 1D-3D en gris et celui des paires 3D-3D en orange.

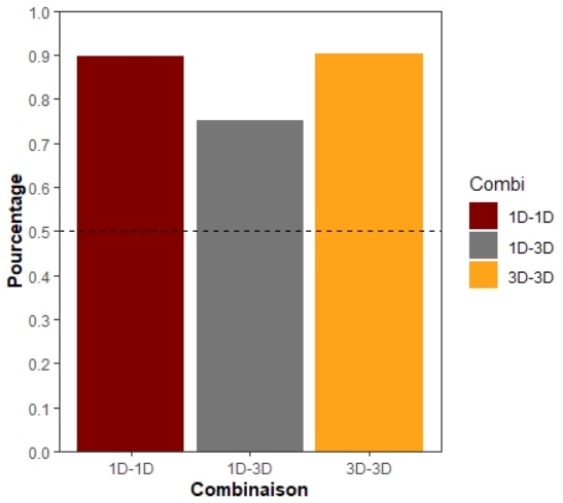

Figure 18 : Pourcentage de réussite des 31 juges à la tâche de discrimination de paires de phonèmes pour chaque type de paire de modèles

D'un point de vue descriptif, nous pouvons voir que les paires 1D-1D et 3D-3D possèdent un taux de discrimination similaire, à presque 90%. Elles sont donc souvent correctement identifiées comme étant identiques. Les paires 1D-3D montrent un taux de discrimination plus faible (74%), indiquant la moindre capacité des juges à détecter la différence. Ce taux reste toutefois supérieur au niveau de hasard (74% > 50%).
D'un point de vue statistique, nous sommes parvenus à montrer un effet significatif du type de paire de modèles avec un X^2 de 76.82, un degré de liberté de 2 et une probabilité inférieure à 0.001. Nous sommes donc en mesure d'**accepter** notre première hypothèse puisque les juges parviennent à percevoir une différence entre les modèles 1D et 3D.

Dans le cas où notre première hypothèse était confirmée, nous avions émis une seconde hypothèse postulant que **les juges seraient aussi performants pour comparer des paires de phonèmes générées à l'aide d'un même modèle acoustique (1D ou 3D) que pour comparer des paires de phonèmes générées à l'aide de deux modèles acoustiques différents (1D et 3D).** La réalisation de contrastes linéaires a permis de situer l'effet significatif relevé avec plus de précision. Les types de paires de modèles ont été comparés deux à deux, c'est-à-dire que les paires 1D-1D ont été comparées aux paires 3D-3D, les paires 1D-1D aux paires 1D-3D et les paires 3D-3D aux paires 1D-3D. Le tableau 3 reprend les données statistiques calculées pour chaque comparaison de paires. Les résultats significatifs ($p \leq 0.05$) sont surlignés en gras.

Type de combinaison du modèle	Score z	Probabilité (> z)
(1D-1D) – (1D-3D)	5.123	**< 0.001**
(3D-3D) – (1D-1D)	0.274	> 0.1
(3D-3D) – (1D-3D)	5.405	**< 0.001**

Tableau 3 : Résultats des contrastes linéaires pour chaque comparaison de paires de modèles

Les résultats révèlent une différence très significative entre les paires 1D-1D et 1D-3D ($z = 5.123$, $p < 0.001$). Une deuxième différence significative a été montrée entre les paires 3D-3D et 1D-3D ($z = 5.405$, $p < 0.001$). En revanche, aucune différence significative n'a pu être démontrée entre les paires 1D-1D et 3D-3D ($z = 0.274$, $p > 0.1$), bien que les paires 3D-3D tendent à être légèrement mieux discriminées que les paires 1D-1D. Notre seconde hypothèse est **infirmée** puisque les juges sont significativement moins performants pour discriminer les paires de phonèmes générées à l'aide de deux modèles acoustiques différents (1D-3D) par

rapport aux paires de phonèmes générées à l'aide d'un même modèle acoustique (1D-1D et 3D-3D).

2.1.1.2. Hypothèse concernant l'effet du genre de la voix de synthèse

Une évaluation de l'effet du genre de la voix de synthèse a eu lieu afin d'observer si le fait d'entendre un phonème généré par une voix féminine ou masculine influence le score de réussite des juges pour discriminer des paires de phonèmes. L'hypothèse énonçait que **les juges auraient de meilleures performances pour comparer des paires de phonèmes générées avec une voix féminine plutôt qu'avec une voix masculine.** Un effet significatif a été obtenu ($X^2 = 10.490$, $p < 0.001$). La figure 19 représente ce résultat. Le pourcentage de réussite de discrimination des paires de phonèmes est présenté sur l'axe des ordonnées. Le genre de la voix de synthèse est indiqué sur l'axe des abscisses, avec 0 « genre féminin » (en rouge) et 1 « genre masculin » (en gris).

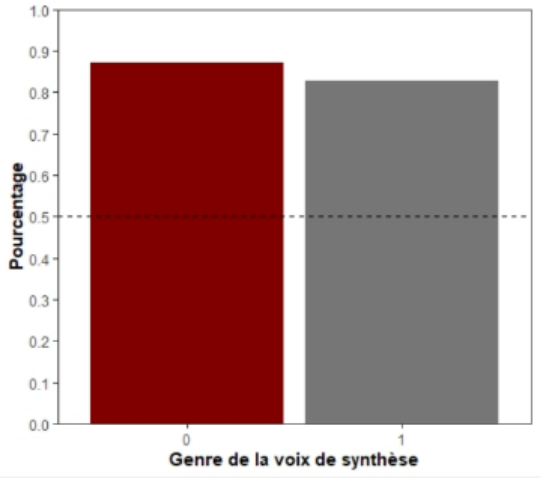

Figure 19 : Pourcentage de réussite des 31 juges à la tâche de discrimination de paires de phonèmes en fonction du genre de la voix de synthèse (avec 0 « genre féminin » et 1 « genre masculin »)

Il est possible de constater que les voix féminines sont un peu mieux discriminées que les voix masculines, à hauteur de 87% (contre 83% pour les voix masculines). Ces résultats **confirment** notre hypothèse**.**

2.1.1.3. Hypothèse concernant la fiabilité intra-juge

Les analyses statistiques n'ont pas permis de montrer un effet significatif du moment ($X^2 = 1.449$, $p > 0.1$). Notre hypothèse selon laquelle **les résultats des juges seront équivalents**

au test et au retest est **confirmée**. Les juges sont fiables car le taux de réussite entre les moments du test et du retest est similaire.

2.1.1.4. Hypothèse concernant la fiabilité inter-juges

Nos analyses statistiques nous ont permis d'obtenir un ICC égal à 0.848. Cet ICC élevé reflète l'important degré d'accord entre les juges. Notre hypothèse, qui suggère que **les réponses des différents juges seront similaires au sein de la tâche de comparaison par paires**, est **confirmée**.

2.1.2. Résultats des hypothèses des effets d'interaction

Outre les effets simples, une hypothèse concernant l'effet d'interaction entre le type de paire de modèles et le type de phonème a été émise. Un éventuel effet permettrait d'apporter une réponse à notre question de recherche qui investigue la manière dont l'emploi de deux modélisations différentes des HF influence la perception de phonèmes. Les effets d'interaction entre le type de paire de modèles et le genre de la voix de synthèse et entre le type de phonème et le genre de la voix de synthèse ont également été investigués. Le tableau 4 expose les résultats obtenus. Lorsque la différence est significative ($p \leq 0.05$), les données sont surlignées en gras. Elles sont détaillées dans les points suivants.

Variables indépendantes	**X^2**	**Degré de liberté (Df)**	**Probabilité (> X^2)**
Type de paire de modèles * type de phonème	97.667	12	**< 0.000**
Type de paire de modèles * genre de la voix de synthèse	3.189	2	> 0.1
Type de phonème * genre de la voix de synthèse	10.052	6	> 0.1

Tableau 4 : Résultats de la régression linéaire mixte généralisée pour les effets d'interaction

2.1.2.1. Hypothèse concernant le type de phonème

Les analyses statistiques de l'effet d'interaction entre le type de paire de modèles et le type de phonème ont été réalisées à l'aide de contrastes linéaires. Chaque type de paire de modèles a été comparée à un autre pour un même phonème (par exemple, [a] 1D-1D comparé

au [a] 1D-3D, [a] 3D-3D comparé au [a] 1D-1D et [a] 3D-3D comparé au [a] 1D-3D). Pour rappel, notre hypothèse indiquait que **les juges seraient plus performants pour comparer des paires de consonnes fricatives non voisées ([s], [ʃ], [f]) générées avec deux modèles acoustiques différents (1D et 3D) que des paires de voyelles ([a], [i] [u] et [ə]) également générées avec deux modèles acoustiques différents (1D et 3D)**. La figure 20 expose le pourcentage de réussite des 31 juges pour discriminer des paires de phonèmes, en tenant compte du phonème et du type de paire de modèles. Le pourcentage de réussite est présenté sur l'axe des ordonnées. Les variables indépendantes du type de phonème et du type de paire de modèles sont indiquées sur l'axe des abscisses.

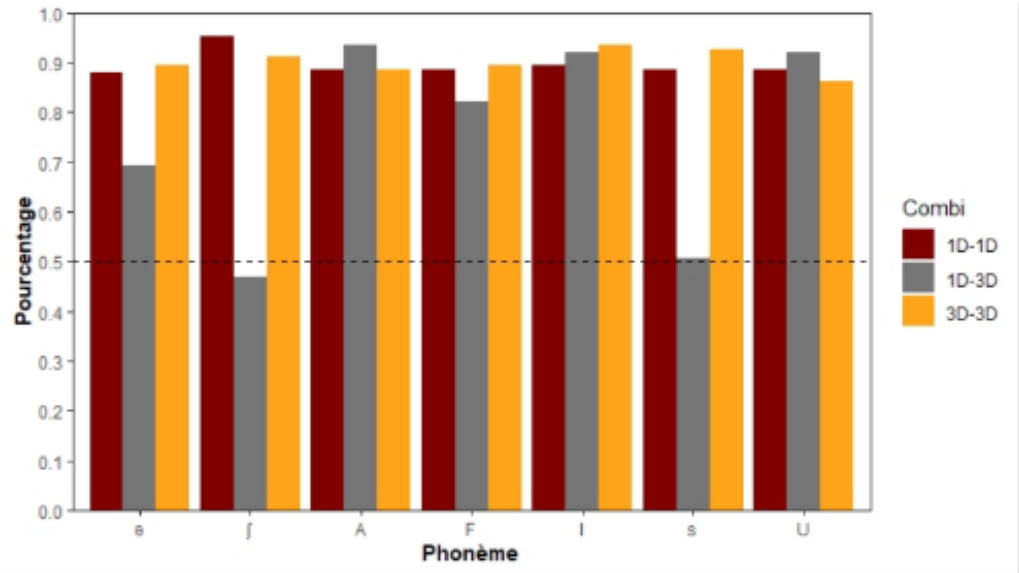

Figure 20 : Pourcentage de réussite des 31 juges pour discriminer des paires de phonèmes, en fonction du phonème et du type de paire de modèles

D'un point de vue descriptif, nous pouvons observer un taux de réussite de presque 90% pour les paires 1D-1D et 3D-3D, peu importe le phonème. Un taux de discrimination 1D-3D beaucoup plus faible est visible pour les phonèmes [ə], [s] et [ʃ]. Le pourcentage de réussite est d'ailleurs à la limite du niveau de hasard pour le [s]. Pour le [ʃ], il est inférieur au seuil de hasard ce qui suggère que les juges ont répondu au hasard à la question qui leur était posée.

Le tableau 5 expose les données statistiques pour l'ensemble des phonèmes testés. Les résultats significatifs ($p \leq 0.05$) sont surlignés en gras. Ils sont discutés à la suite du tableau.

Phonème	Type d'appariement de modèles acoustiques	Score z	Probabilité (> z)
[ə]	(1D-1D) – (1D-3D)	3.502	**< 0.1**
	(3D-3D) – (1D-1D)	0.402	> 0.1
	(3D-3D) – (1D-3D)	3.816	**< 0.05**
[ʃ]	(1D-1D) – (1D-3D)	7.000	**< 0.001**
	(1D-1D) – (3D-3D)	1.210	> 0.1
	(3D-3D) – (1D-3D)	6.947	**< 0.001**
[a]	(1D-3D) – (1D-1D)	1.437	> 0.1
	(1D-1D) – (3D-3D)	0.066	> 0.1
	(1D-3D) – (3D-3D)	1.493	> 0.1
[f]	(1D-1D) – (1D-3D)	1.347	> 0.1
	(3D-3D) – (1D-1D)	0.125	> 0.1
	(3D-3D) – (1D-3D)	1.462	> 0.1
[i]	(1D-3D) – (1D-1D)	0.740	> 0.1
	(3D-3D) – (1D-1D)	1.104	> 0.1
	(3D-3D) – (1D-3D)	0.381	> 0.1
[s]	(1D-1D) – (1D-3D)	6.181	**< 0.001**
	(3D-3D) – (1D-1D)	1.036	> 0.1
	(3D-3D) – (1D-3D)	6.581	**< 0.001**
[u]	(1D-3D) – (1D-1D)	0.882	> 0.1
	(1D-1D) – (3D-3D)	0.564	> 0.1
	(1D-3D) – (3D-3D)	1.426	> 0.1

Tableau 5 : Résultats des contrastes linéaires pour chaque type de paire de modèles pour l'ensemble des phonèmes

L'analyse des phonèmes [a], [f], [i] et [u] ne permet pas de mettre en évidence de différence significative entre les dimensions des modèles acoustiques appariés. Le [ə], le [s] et le [ʃ] présentent des résultats différents. Bien que les paires 1D-1D ne soient jamais mieux discriminées que les paires 3D-3D et inversement, une différence significative est presque systématiquement relevée lorsque la combinaison 1D-3D est considérée. Pour le [ə], les paires 1D-1D sont légèrement mieux discriminées que les paires 1D-3D ($z = 3.502$, $p < 0.1$) mais la différence n'est pas significative. Un effet significatif est toutefois observé pour les paires 3D-3D qui sont mieux discriminées que les paires 1D-3D ($z = 3.816$, $p < 0.05$). Pour le [ʃ], les

paires 1D-1D ($z = 7.000$, $p < 0.001$) et 3D-3D ($z = 6.947$, $p < 0.001$) sont nettement mieux discriminées que les paires 1D-3D. Pour le [s], les paires 1D-1D ($z = 6.181$, $p < 0.001$) et 3D-3D ($z = 6.581$, $p < 0.001$) sont nettement mieux discriminées que les paires 1D-3D.
Un effet simple significatif lié au phonème a été mis en avant par la régression linéaire mixte généralisée ($X^2 = 46.62$, $p < 0.000$) (voir tableau 2). Lorsque l'on compare les scores de réussite des phonèmes entre eux, sans tenir compte du type de paire de modèles, trois effets significatifs sont mis en évidence. Le [ə] est significativement moins bien discriminé que le [i] ($z = 3.294$, $p < 0.05$). Le [s] est significativement moins bien discriminé que le [a] ($z = 3.082$, $p < 0.05$) et le [i] ($z = 3.538$, $p < 0.01$). Au regard de ces résultats, nous **ne pouvons pas confirmer** notre hypothèse puisque la paire 1D-3D du [f] n'est pas mieux discriminée que celle des voyelles [a], [i], [u] et [ə] et que la paire 1D-3D du [s] et du [ʃ] est significativement moins bien discriminée par rapport aux autres phonèmes.

2.1.2.2. Analyse complémentaire concernant l'effet d'interaction entre le type de paire de modèles et le genre de la voix de synthèse

Les analyses statistiques à propos de l'effet d'interaction entre le type de paire de modèles et le genre de la voix de synthèse n'ont pas permis de mettre en évidence un effet significatif entre les deux variables ($z = 3.189$, $p > 0.1$). L'association d'une paire de modèles spécifique avec le genre vocal masculin ou féminin ne semble pas influencer la capacité de discrimination des juges. Ces analyses étant complémentaires, nous n'avions pas émis d'hypothèse à leur propos. Elles ne seront donc pas traitées dans la discussion.

2.1.2.3. Analyse complémentaire concernant l'effet d'interaction entre le type de phonème et le genre de la voix de synthèse

Nous avons souhaité déterminer si l'association entre un phonème particulier et le genre de la voix de synthèse influence la capacité des juges à discriminer des paires de phonèmes. Nos analyses n'ont pas permis de montrer un effet significatif concernant cette interaction ($z = 10.052$, $p > 0.1$). La capacité de discrimination des juges ne paraît pas être influencée par l'association entre le type de phonème et le genre vocal masculin ou féminin. Ces résultats sont complémentaires. Ils ne seront pas abordés dans la section « discussion » car aucune hypothèse n'avait été énoncée.

2.2. Résultats des hypothèses de la seconde tâche expérimentale

2.2.1. Résultats des hypothèses des effets simples

Les résultats statistiques des hypothèses des effets simples sont exposés dans le tableau 6. Il reprend les probabilités calculées pour chacune des variables mentionnées précédemment. Lorsque l'effet est significatif ($p \leq 0.05$), le résultat est surligné en gras. Chacun des effets et ses résultats sont discutés séparément dans les points 2.2.2.1., 2.2.2.2., 2.2.2.3. et 2.2.2.4..

Variable indépendante	Rapport de vraisemblance (LR)	Degré de liberté (Df)	Probabilité (> X^2)
Réalisme du modèle acoustique	2.961	1	< 0.1
Genre de la voix de synthèse	1.130	1	> 0.1
Type de phonème	464.038	6	**< 0.001**
Moment	0.518	1	> 0.1

Tableau 6 : Résultats du test du rapport de vraisemblance des modèles au niveau des effets simples

2.2.1.1. Hypothèse concernant l'effet du réalisme du modèle acoustique

Dans un premier temps, nous avons évalué l'influence du modèle 1D ou 3D sur le score attribué à l'aspect naturel des phonèmes. Pour rappel, notre hypothèse stipulait que **les phonèmes générés par le modèle 3D seraient perçus comme plus naturels que ceux générés par le modèle 1D**. Nous ne sommes pas parvenus à montrer un effet significatif de la modélisation utilisée sur la perception du degré de naturel des phonèmes ($LR = 2.961$, $p > 0.1$). Notre hypothèse **ne peut donc pas être confirmée**. Compte-tenu l'absence de significativité de l'effet, nous n'avons pas réalisé de contrastes linéaires pour approfondir les résultats.

2.2.1.2. Hypothèse concernant l'effet du genre de la voix de synthèse

Nous avons investigué un éventuel effet du genre de la voix de synthèse sur le score attribué à l'aspect naturel des phonèmes par les juges. Aucune différence significative n'a pu être démontrée ($LR = 1.130$, $p > 0.1$). Nous **ne sommes pas en mesure de confirmer** notre hypothèse selon laquelle **les phonèmes générés avec une voix féminine et ceux générés avec une voix masculine devraient être perçus avec le même degré de naturel**. La non-significativité de l'effet ne nous a pas permis de réaliser des contrastes linéaires pour investiguer ce résultat davantage.

2.2.1.4. Hypothèse concernant la fiabilité intra-juge

Les analyses statistiques n'ont pas permis de mettre en évidence de différence significative concernant l'effet du moment de présentation du stimulus durant le testing (LR = 0.518, p > 0.1). L'aspect naturel des phonèmes a été évalué de la même manière entre le moment du test et celui du retest. Notre hypothèse est donc **confirmée** puisque **les résultats des juges sont équivalents au test et au retest**. Aucun contraste linéaire n'a été réalisé au regard de la non-significativité de l'effet.

2.2.1.5. Hypothèse concernant la fiabilité inter-juges

Un ICC égal à 0.919 a été obtenu. Ce résultat illustre un degré d'accord important entre les juges. Les juges étant fiables entre eux, nous pouvons **confirmer notre hypothèse** selon laquelle **les résultats entre les juges sont équivalents concernant le score de degré de naturel attribué aux phonèmes**.

2.2.2. Résultats des hypothèses des effets d'interaction

Plusieurs effets d'interaction ont été investigués dans cette seconde tâche expérimentale. Celui entre le réalisme du modèle acoustique et le type de phonème est présenté dans le point 2.2.2.1., suivi de celui entre le réalisme du modèle acoustique et le genre de la voix de synthèse dans le point 2.2.2.2.. Le tableau 7 expose les résultats obtenus.

Variables indépendantes	Rapport de vraisemblance	Degré de liberté (Df)	Probabilité (> X^2)
Réalisme du modèle acoustique * type de phonème	6.821	6	> 0.1
Réalisme du modèle acoustique * genre de la voix de synthèse	0.021	1	> 0.1

Tableau 7 : Résultats du test du rapport de vraisemblance des modèles au niveau des effets d'interaction

2.2.2.1. Hypothèse concernant l'effet du type de phonème

Le test du rapport de vraisemblance des modèles de liens cumulatifs a mis en évidence une différence significative lorsque la variable indépendante « type de phonème » est introduite (LR = 464.038, p < 0.001) (voir tableau 6). Notre hypothèse selon laquelle **le score attribué**

pour le degré de naturel varie selon les phonèmes est **confirmée**. Nous nous attendions plus particulièrement à ce que **les consonnes fricatives non voisées [s], [ʃ], [f], générées avec le modèle 3D, soient considérées comme plus naturelles que les voyelles [a], [i] [u] et [ə], générées avec le modèle 3D**. Afin de répondre à cette hypothèse, nous avons d'abord investigué l'effet simple du type de phonème. Puis, nous nous sommes intéressés à l'effet d'interaction entre le réalisme du modèle acoustique et le type de phonème.

Dans un premier temps, des contrastes linéaires entre les phonèmes ont été réalisés pour approfondir la significativité de l'effet du type de phonème. Autrement dit, le score moyen de l'aspect naturel de chaque phonème a été comparé à celui des six autres phonèmes. La figure 21 représente le score moyen attribué par les 31 juges pour l'aspect naturel de chaque phonème évalué.

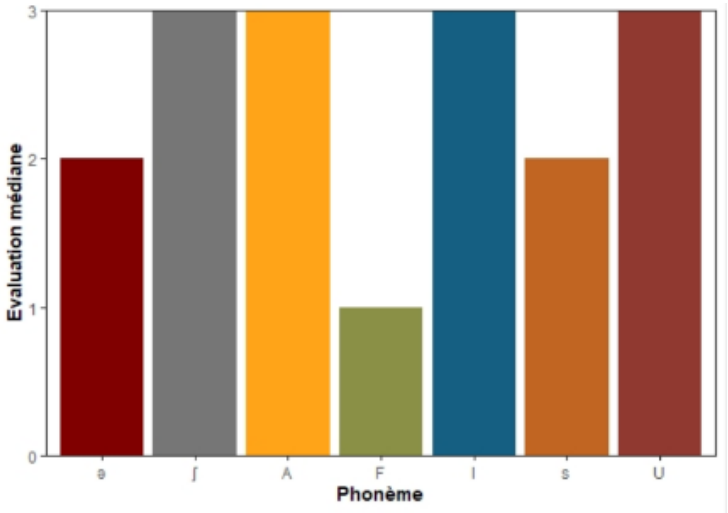

Figure 21 : Score moyen attribué à l'aspect naturel de chaque phonème

D'un point de vue descriptif, il est possible de constater que certains phonèmes sont considérés comme totalement naturels tandis que d'autres le sont beaucoup moins. Les phonèmes [ʃ], [a], [i] et [u] sont perçus avec un haut degré de naturel, contrairement aux phonèmes [ə], [f] et [s]. D'un point de vue statistique, les contrastes linéaires ont permis de relever plusieurs différences significatives. Les résultats sont exposés dans le tableau 8 ci-dessous. Ils sont surlignés en gras lorsque la différence est significative ($p \leq 0.05$).

Phonème	Phonème contrasté	Score z	Probabilité (> z)
[ə]	[ʃ]	2.014	**< 0.05**
	[a]	4.717	**< 0.001**
	[f]	-9.038	**< 0.001**

	[i]	3.137	**< 0.01**
	[s]	-0.235	> 0.1
	[u]	0.170	> 0.1
[ʃ]	[a]	2.703	**< 0.01**
	[f]	-10.554	**< 0.001**
	[i]	1.117	> 0.1
	[s]	-2.215	**< 0.05**
	[u]	-1.852	< 0.1
[a]	[f]	-12.666	**< 0.001**
	[i]	-1.594	> 0.1
	[s]	-4.874	**< 0.001**
	[u]	-4.566	**< 0.001**
[f]	[i]	11.455	**< 0.001**
	[s]	8.750	**< 0.001**
	[u]	9.189	**< 0.001**
[i]	[s]	-3.321	**< 0.001**
	[u]	-2.980	**< 0.01**
[s]	[u]	0.403	> 0.1

Tableau 8 : Résultats des contrastes linéaires au niveau du phonème

Lorsque nous considérons le score d'aspect naturel du [ə], nous observons qu'il est significativement différent de celui du [ʃ] ($z = 2.014$, $p < 0.05$), du [a] ($z = 4.717$, $p < 0.001$), du [f] ($z = -9.038$, $p < 0.001$) et du [i] ($z = 3.137$, $p < 0.01$). Aucune différence significative n'a pu être mise en évidence avec le [s] (-0.235, $p > 0.1$) et le [u] ($z = 0.170$, $p > 0.1$).

En ce qui concerne le [ʃ], son score est significativement différent de celui du [a] ($z = 2.703$, $p < 0.01$), du [f] ($z = -10.554$, $p < 0.001$) et du [s] ($z = -2.215$, $p < 0.05$), en plus du [ə] déjà mentionné ($z = -2.014$, $p < 0.05$). Il n'est pas significativement différent du [i] ($z = 1.117$, $p > 0.1$) et du [u] ($z = -1.852$, $p < 0.1$).

Outre les différences significatives avec le [ə] ($z = -4.717$, $p < 0.001$) et le [ʃ] (-2.703, $p < 0.01$), le [a] présente un score d'aspect naturel significativement différent de celui du [f] ($z = -12.666$, $p < 0.001$), du [s] ($z = -4.874$, $p < 0.001$) et du [u] ($z = -4.566$, $p < 0.001$). Nous ne sommes pas parvenus à montrer de différence significative avec le [i] ($z = -1.594$, $p > 0.1$).

Le score du [f] diffère significativement de l'ensemble des phonèmes testés dans cette étude, avec une probabilité toujours inférieure à 0.001.

A propos du score du [i], nous avons observé une différence significative avec celui du [s] ($z = -3.321$, $p < 0.001$) et du [u] ($z = -2.980$, $p < 0.001$), en plus de celle avec le [ə] ($z = -3.137$, $p < 0.01$) et le [f] ($z = -11.455$, $p < 0.001$). Comme indiqué précédemment, aucune différence significative n'a pu être démontrée entre le [i] et les phonèmes [ʃ] ($z = -1.116$, $p > 0.1$) et [a] ($z = 1.594$, $p > 0.1$).

Nous avons déjà montré les différences significatives entre le score du [s] et celui du [ʃ] ($z = 2.215$, $p < 0.05$), du [a] ($z = 4.874$, $p < 0.001$), du [f] ($z = -8.751$, $p < 0.001$) et du [i] ($z = 3.321$, $p < 0.001$). Le contraste linéaire réalisé entre le score du [s] et celui du [u] ne nous a pas permis de montrer de différence significative ($z = 0.403$, $p > 0.1$).

Lorsque nous nous intéressons au score du [u], nous relevons une différence significative avec celui du [a] ($z = 4.566$, $p < 0.001$), du [f] ($z = -9.190$, $p < 0.001$) et du [i] ($z = 2.980$, $p < 0.01$). Nous ne sommes pas parvenus à mettre en évidence une différence significative avec le [ə] ($z = -0.170$, $p > 0.1$), le [ʃ] ($z = 1.852$, $p < 0.1$) et le [s] ($z = -0.403$, > 0.1).

Nous avons ensuite investigué l'effet d'interaction entre le réalisme du modèle acoustique et le type de phonème afin de répondre à notre hypothèse. La figure 22 représente les scores de degré de naturel obtenus pour chaque phonème, en fonction des modélisations 1D et 3D.

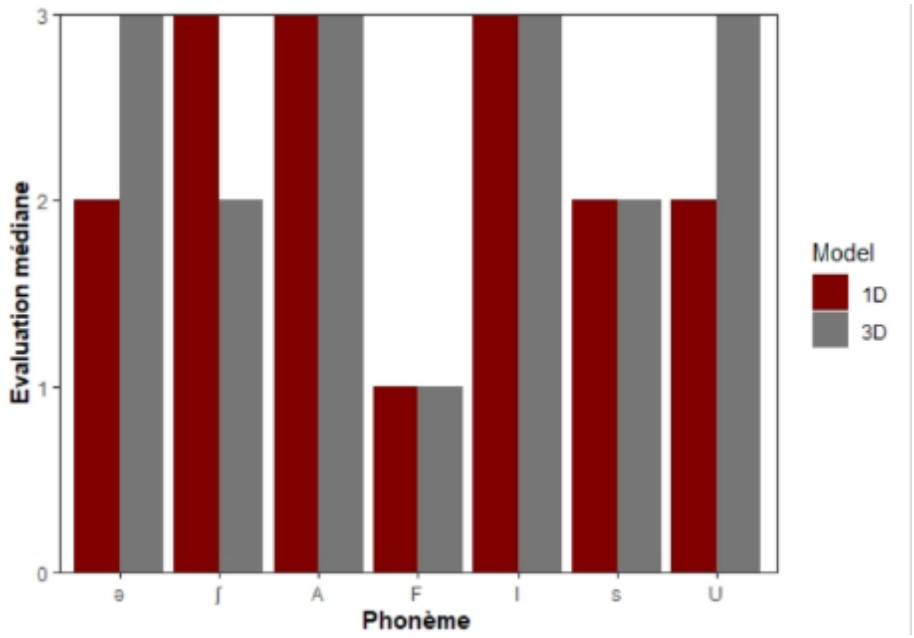

Figure 22 : Score moyen d'évaluation de l'aspect naturel de chaque phonème, en fonction des modélisations 1D et 3D

D’un point de vue descriptif, nous pouvons observer que le [a] et le [i] sont considérés comme « totalement naturels », peu importe la modélisation utilisée. Le [ə] 3D, le [ʃ] 1D et le [u] 3D sont également perçus comme « totalement naturels ». Le [f] n’est pas considéré comme naturel, qu’il soit généré avec le modèle 1D ou 3D. Enfin, le [ə] 1D, le [s] 1D et 3D et le [u] 1D sont évalués comme moins naturels par rapport aux autres.

Les analyses statistiques n’ont pas permis de montrer de différence significative, pour aucun des phonèmes ($LR = 6.821$, $p < 0.1$). Notre hypothèse, selon laquelle **les consonnes fricatives non voisées [s], [ʃ], [f], générées avec le modèle 3D, seront considérées comme plus naturelles que les voyelles [a], [i] [u] et [ə], générées avec le modèle 3D**, **ne peut pas être confirmée**. En raison de l’absence de significativité du résultat, nous n’avons pas pu réaliser de contrastes linéaires au niveau de l’interaction entre le réalisme du modèle acoustique et le type de phonème.

2.2.2.2. Analyse complémentaire concernant l’effet d’interaction entre le réalisme du modèle acoustique et le genre de la voix de synthèse

Nous nous sommes intéressés à déterminer si l’association entre le réalisme du modèle acoustique et le genre de la voix de synthèse influence l’aspect naturel des phonèmes. Nous ne sommes pas parvenus à montrer de différence significative concernant cet effet d’interaction ($LR = 0.021$, $p > 0.1$). Peu importe qu’il s’agisse d’une voix synthétique féminine ou masculine, le score d’aspect naturel attribué aux phonèmes 3D ne diffère pas significativement de celui des phonèmes 1D. Compte-tenu l’absence de significativité du résultat, nous n’avons pas réalisé de contrastes linéaires. Cette analyse étant complémentaire, nous n’avions pas émis d’hypothèse à son propos. Elle ne sera donc pas traitée dans la discussion.

VI. DISCUSSION

1. Rappel des objectifs de l'étude

Pour rappel, ce mémoire s'inscrit dans un projet de développement d'une synthèse articulatoire à large bande. Sa visée est de mieux comprendre et définir le lien entre la perception de la parole et les aspects physiques et acoustiques liés à sa production dans l'entièreté du registre fréquentiel audible (de 0.02 à 20 kHz). Afin de répondre à cette problématique, nous avons cherché à objectiver la manière dont l'emploi de deux modélisations différentes des HF influence la perception de phonèmes. Deux tâches perceptives ont été élaborées autour de phonèmes pour lesquels la modélisation physique des HF était différente. Celles-ci ont été générées à partir de deux modèles différents : le modèle 1D et le modèle 3D. Dans un premier temps, la sensibilité auditive de jeunes adultes à ces deux degrés de réalisme dans la synthèse des HF a été évaluée. Dans un second temps, nous avons souhaité déterminer si une description correcte de la physique dans la synthèse des HF, telle que dans le modèle 3D, contribue à l'aspect naturel de la parole artificielle.
Plusieurs hypothèses ont été émises pour chacune des tâches perceptives à propos du modèle acoustique utilisé, du genre de la voix de synthèse, du type de phonème et des fiabilités intra- et inter-juges. Des effets d'interaction ont également été investigués entre le modèle acoustique et le type de phonème. Les analyses statistiques complémentaires présentées dans la section « résultats » ne sont pas abordées dans cette discussion car nous n'avions pas émis d'hypothèse à leur propos.

Cette discussion est composée de deux parties. D'une part, les réponses aux hypothèses sont exposées et interprétées. D'autre part, les limites et biais méthodologiques sont discutés afin de proposer des perspectives d'amélioration pour ce type d'étude.

2. Réponses aux hypothèses et interprétation des résultats

2.1. Résultats et analyses de la première expérience

Notre première expérience visait à évaluer la capacité des juges à discriminer des paires de phonèmes, pour lesquels les HF étaient modélisées de deux manières différentes. Les résultats des hypothèses liées au type de paire de modèles, au genre de la voix de synthèse, au type de phonème et aux fiabilités intra- et inter-juges sont respectivement discutés dans les points 2.1.1., 2.1.2., 2.1.3., 2.1.4. et 2.1.5..

2.1.1. Le type de paire de modèles

Nous avons cherché à déterminer si le type de modélisation des HF (1D ou 3D) influence la faculté de discrimination des juges pour des paires de phonèmes. Pour ce faire, 4 voyelles ([a], [i], [u], [ə]) et 3 consonnes fricatives non voisées ([f], [s], [ʃ]) ont été générées à l'aide de la synthèse articulatoire. Pour rappel, les phonèmes ont été amputés de leur énergie dans les basses fréquences afin de nous assurer que seules les HF étaient évaluées par les juges (Vitela et al., 2015). Trois types de paires de modèles ont été introduits : les paires 1D-1D, les paires 3D-3D et les paires 1D-3D.

Selon plusieurs auteurs (Monson et al., 2011 ; Monson & Caravello, 2019), le système auditif humain est capable de percevoir les HF dans un signal de parole. Les humains devraient donc être capables d'entendre des différences au sein de la gamme des HF dans la parole. Une première hypothèse a été formulée. Elle stipule que **des différences seraient perçues entre les phonèmes générés par le modèle 1D et ceux générés par le modèle 3D**. Les résultats ont montré un pourcentage de discrimination correcte des paires 1D-3D supérieur au niveau de hasard (74% > 50%). Par ailleurs, un effet significatif du type de paire de modèles a été mis en évidence ($X^2 = 76.82$, $p < 0.001$). Notre première hypothèse est **confirmée** puisque les juges parviennent à percevoir une différence entre les modèles 1D et 3D.

Pour rappel, une audiométrie tonale avait été réalisée afin de tester les seuils auditifs des juges dans les HF, à 12.5 kHz et 16 kHz. Aucun juge n'avait présenté de mauvais seuils et été exclu de l'étude (voir annexe 4). Leur sensibilité auditive aux HF étant bonne, nous avions émis l'hypothèse que **les juges seraient aussi performants pour comparer des paires de phonèmes générées à l'aide d'un même modèle acoustique (1D ou 3D) que pour comparer des paires de phonèmes générées à l'aide de deux modèles acoustiques différents (1D et 3D).** En d'autres termes, nous nous attendions à ce que les scores de discrimination des paires 1D-1D, 3D-3D et 1D-3D soient équivalents. Une différence significative concernant le pourcentage de discrimination correcte a été relevée entre les paires 1D-1D et 1D-3D ($z = 5.123$, $p < 0.001$) et les paires 3D-3D et 1D-3D ($z = 5.405$, $p < 0.001$). Notre hypothèse **ne peut pas être confirmée** puisque les scores ne sont pas équivalents. La différence entre les paires 1D-3D est moins bien perçue. Les paires de modèles acoustiques hétérogènes (1D-3D) sont plus faiblement discriminées que les paires de modèles acoustiques homogènes (1D-1D et 3D-3D).

Il convient de préciser que ces résultats restent exploratoires. Aucune étude à l'heure actuelle ne semble avoir évalué les différences perceptives générées par l'usage de modèle 1D

et 3D auprès d'auditeurs humains. Il est donc difficile d'offrir une interprétation au regard de ce qui a été révélé dans la littérature scientifique et de la dimension novatrice de notre étude.

2.1.2. Le genre de la voix de synthèse

Concernant le genre de la voix de synthèse, nous nous attentions à ce que **les juges aient de meilleures performances pour comparer des paires de phonèmes générées avec une voix féminine plutôt qu'avec une voix masculine**. Un effet significatif a été mis en évidence ($X^2 = 10.49$, $p < 0.001$) qui **confirme** notre hypothèse. Les paires de voix féminines ont été mieux discriminées que les paires de voix masculines, avec 87% de discrimination correcte contre 83%. Plusieurs auteurs ont mis en avant l'influence du genre du locuteur sur la capacité de perception des HF dans la parole (Monson et al., 2012 ; Stelmachowitcz et al., 2001). Cette influence peut s'expliquer par le fait que les bandes d'octave et de tiers d'octave à 8 kHz, 10.1 kHz, 12.7 kHz, 16 kHz et 20.2 kHz sont plus riches en HF dans les voix féminines (Monson et al., 2012). Les juges n'ayant entendu que les fréquences supérieures à 5 kHz des phonèmes en raison du filtrage de leurs basses fréquences, les différences de niveaux de HF étaient probablement davantage marquées dans les paires de voix féminines.

2.1.3. Le type de phonème

Les consonnes fricatives non voisées [s], [ʃ] et [f] étant plus riches en HF par rapport aux voyelles (Monson, Hunter et al., 2014), nous avons supposé que leur distinction selon les modèles 1D et 3D devrait être plus simple. En effet, si l'intensité des HF est plus élevée, la différence entre un phonème 1D et un phonème 3D devrait être plus marquée. Nous avons donc émis l'hypothèse que **les juges seraient plus performants pour comparer des paires de consonnes fricatives non voisées ([s], [ʃ], [f]) générées avec deux modèles acoustiques différents (1D et 3D) que des paires de voyelles ([a], [i] [u] et [ə]) également générées avec deux modèles acoustiques différents (1D et 3D)**. Autrement dit, le score de discrimination des paires 1D-3D devrait être supérieur pour les consonnes fricatives non voisées par rapport aux voyelles. Nous ne sommes pas parvenus à montrer de différence significative entre les types de paires de modèles pour les phonèmes [a], [f], [i] et [u]. Les paires 1D-1D ne sont donc pas mieux discriminées que les paires 1D-3D, les paires 3D-3D ne sont pas mieux discriminées que les paires 1D-1D et les paires 3D-3D ne sont pas mieux discriminées que les paires 1D-3D. Ces constats sont valables pour les quatre phonèmes mentionnés. Des effets significatifs ont en revanche été relevés pour le [ə], le [s] et le [ʃ], uniquement lorsque l'appariement 1D-3D était pris en compte. Les paires 3D-3D sont significativement mieux discriminées que les paires 1D-

3D pour le [ə] ($z = 3.816$, $p < 0.05$). La discrimination des paires 1D-1D du [ə] tend à être meilleure que les paires 1D-3D mais l'effet n'est pas significatif. Concernant le [s], les paires 1D-3D sont nettement moins bien discriminées que les paires homogènes 1D-1D ($z = 6.181$, $p < 0.001$) et 3D-3D ($z = 6.581$, $p < 0.001$). Il en est de même pour le [ʃ], pour lequel les paires 1D-3D sont nettement moins bien discriminées que les paires 1D-1D ($z = 7.000$, $p < 0.001$) et 3D-3D ($z = 6.947$, $p < 0.001$). Les phonèmes [ə], [s] et [ʃ] engendrent ainsi une diminution très forte de la capacité de discrimination des juges au sein des paires 1D-3D par rapport aux paires 1D-1D et 3D-3D. Notre hypothèse **ne peut pas être confirmée** car la paire 1D-3D du [f] n'est pas mieux discriminée que celle des voyelles [a], [i], [u] et [ə] et la paire 1D-3D du [s] et du [ʃ] est significativement moins bien discriminée par rapport aux autres phonèmes. Au contraire, les phonèmes les mieux discriminés avec les paires 1D-3D sont les voyelles [a], [i] et [u].

En outre, un effet significatif du type de phonème, sans tenir compte des paires de modèles, a été mis en évidence ($X^2 = 46.62$, $p < 0.000$). Il se peut donc que les résultats des phonèmes [ə], [s] et [ʃ] ne s'expliquent pas en termes de type de paire de modèles. En effet, les phonèmes seuls n'ont pas tous le même impact concernant la perception de différences entre les paires de sons. Nous nous sommes interrogés sur l'origine de l'effet simple du type de phonème afin d'expliquer les résultats de l'effet d'interaction.

Une interprétation possible des résultats concerne les différences de niveaux des HF en décibels entre les phonèmes 1D et 3D. La figure 23 représente ces différences de niveaux. L'axe des abscisses expose chaque phonème, selon les genres féminin (en rouge) et masculin (en gris). L'axe des ordonnées indique le niveau en HF en décibels. Ce graphique a été obtenu à partir d'une soustraction entre le niveau des HF (en dB) du phonème généré par le modèle 3D et le niveau des HF (en dB) du même phonème généré par le modèle 1D. Lorsque la différence est positive, le phonème 1D est plus riche en HF que le phonème 3D. Lorsque la différence est négative, le phonème 3D est plus riche en HF que le phonème 1D. Nous considérons qu'à partir de 1 dB de différence du niveau de pression sonore, celle-ci est assez importante pour être perceptivement audible (R. Blandin, communication personnelle, 30 juin 2021).

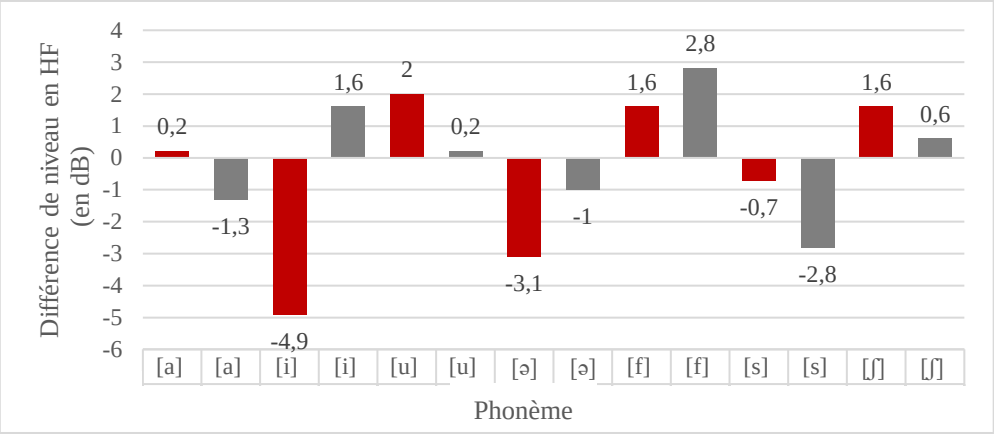

Figure 23 : Différences de niveaux en HF en décibels pour chaque phonème, selon les genres féminin (en rouge) et masculin (en gris)

D'un point de vue descriptif, il est possible de constater que le niveau en HF semble plutôt aléatoire d'un phonème à l'autre. Pour le [u], le [f] et le [ʃ], les différences sont positives, peu importe le genre de la voix de synthèse. Ceci indique que leur modélisation 1D est plus riche en HF que leur modélisation 3D. Pour le [ə] et le [s], les différences sont négatives, quel que soit le genre de la voix de synthèse. Contrairement au [u], au [f] et au [ʃ], leur modélisation 3D est plus riche en HF. Les observations sont différentes pour les phonèmes [a] et [i], qui présentent une différence tantôt positive, tantôt négative, selon qu'il s'agisse d'un genre vocal féminin ou masculin. Outre la valence de la différence, le niveau en HF est disparate selon le phonème considéré. Lorsque nous représentons ces données sur une droite de régression, le caractère aléatoire est visible (voir figure 24). Chaque point représente un phonème selon le genre féminin (en rouge) ou masculin (en gris).

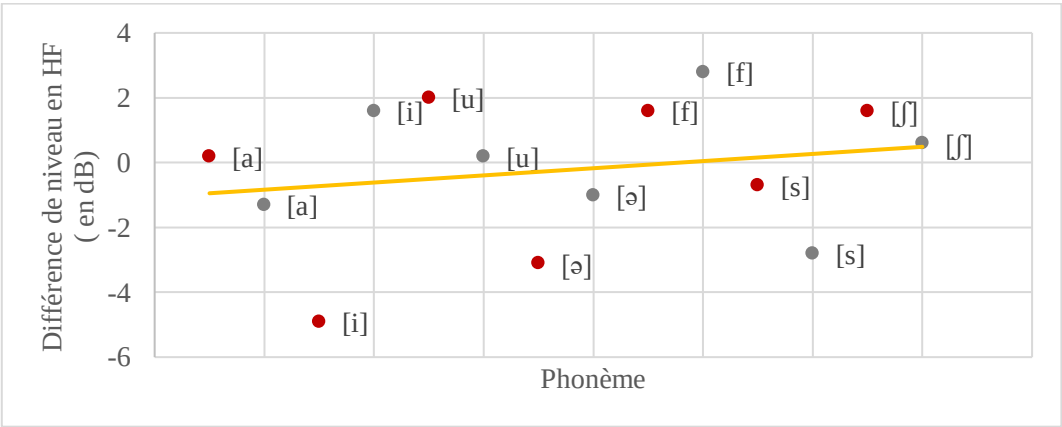

Figure 24 : Droite de régression représentant les différences de niveaux en HF (en dB) entre les modèles 1D et 3D, pour chaque phonème selon les genres féminin (en rouge) et masculin (en gris)

Nous pouvons voir que les points sont fortement dispersés autour de la droite. Plus précisément, les phonèmes [ə], [s] et [ʃ] ne semblent pas montrer de tendance particulière. Le caractère aléatoire exclut ainsi l'interprétation des résultats par les différences de niveaux dans les HF. Il conviendrait néanmoins d'objectiver cette interprétation subjective par une analyse statistique afin de statuer sur la présence ou l'absence de différence significative.

2.1.4. La fiabilité intra-juge

Evaluer l'effet du moment de présentation de paires de phonèmes pendant le testing nous a permis de nous assurer de la fiabilité intra-juge de chaque participant. Pour rappel, elle correspond au degré d'accord du juge envers lui-même quant à la réponse donnée. Elle a été vérifiée par la double présentation de chaque paire de phonèmes, à deux moments différents du testing. Ceux-ci sont référés par les termes « test » et « retest ». L'hypothèse selon laquelle **les résultats des juges seraient équivalents entre le test et le retest** avait été émise. **Conformément à notre hypothèse**, aucune différence significative n'a pu être montrée concernant le moment de présentation des paires de phonèmes ($X^2 = 1.449$, $p > 0.1$). En d'autres termes, les juges ont aussi bien réussi à discriminer les paires de sons lors du test que lors du retest. Ce résultat concorde avec les dires de Teston (2004), qui affirme qu'une tâche de comparaison par paires augmente la fiabilité intra-juge lors d'une tâche perceptive, en particulier quand les paires à évaluer s'enchaînent sans pause. Par ailleurs, ces données suggèrent qu'il n'y a pas eu d'effet d'apprentissage au cours de l'expérience. Les juges n'ont pas appris à discriminer les paires de phonèmes durant la phase de test, auquel cas ils auraient obtenu de meilleurs scores durant le retest.

2.1.5. La fiabilité inter-juges

Nous avons souhaité évaluer la fiabilité inter-juges afin de vérifier l'homogénéité des réponses données par l'ensemble de la cohorte. Autrement dit, nous nous sommes intéressés au degré de consensus des juges quant au fait de déterminer si la paire de phonèmes était identique ou différente. L'hypothèse selon laquelle **les réponses des différents juges seraient similaires au sein de la tâche de comparaison par paires** a été émise. Afin d'y répondre, un effet aléatoire lié aux juges a été introduit dans notre modèle statistique. L'ICC élevé, égal à 0.848, montre que les juges n'ont pas évalué les paires de phonèmes différemment. D'après Koo et Li (2016), nous pouvons considérer que la fiabilité inter-juges est bonne. Ils sont donc considérés comme fiables, ce qui nous permet de **confirmer** notre hypothèse.

2.2. Résultats et analyses de la seconde expérience

La seconde expérience visait à évaluer dans quelle mesure la modélisation 3D des HF est susceptible d'influencer la perception du degré de naturel d'un signal de parole, en comparaison à la modélisation 1D. Pour ce faire, une tâche de jugement de l'aspect naturel de phonèmes isolés a été proposée aux juges. Les points suivants exposent les résultats et leur interprétation concernant les hypothèses liées au réalisme du modèle acoustique, au genre de la voix de synthèse, au type de phonème et aux fiabilités intra- et inter-juges.

2.2.1. Réalisme du modèle acoustique

Nous avons supposé que **les phonèmes générés par le modèle 3D seraient perçus comme plus naturels que ceux générés par le modèle 1D**. Les analyses statistiques n'ont pas permis de mettre en évidence de différence significative en fonction du modèle utilisé ($LR = 2.961$, $p > 0.1$). **Contrairement à notre hypothèse**, les phonèmes générés par le modèle 1D et ceux générés par le modèle 3D ont été perçus avec un degré de naturel équivalent.

La non-significativité du résultat peut éventuellement être expliquée par l'emploi de phonèmes isolés monotones. Leur aspect naturel est probablement difficile à évaluer en raison de l'absence de parole connectée, qui prive nos phonèmes de variations intonatives. D'après Anand et Stepp (2015), un signal de parole monotone tend à être associé à un degré de naturel plus faible. L'objectif des auteurs était d'investiguer le lien entre la perception de la monotonie et l'aspect naturel de la parole. La monotonie réfère à l'absence de variation de hauteur et d'inflexion dans la voix. Un total de 16 auditeurs non experts (6 hommes et 10 femmes), âgés de 18 à 27 ans, devaient évaluer la monotonie, l'aspect naturel et l'intelligibilité de deux phrases, chacune produite par un patient parkinsonien dont la sévérité de la dysarthrie était de légère à sévère. Une forte corrélation entre la monotonie et l'aspect naturel des signaux de parole connectée a été montrée. Dans le cadre de notre étude, la monotonie des phonèmes est donc susceptible d'avoir influencé l'évaluation des juges pour leur aspect naturel.

En outre, nous sommes peu amenés à écouter des phonèmes isolés dans notre quotidien. Employer de la parole connectée est donc une piste pour mettre en évidence des différences entre les modèles 1D et 3D à propos de l'aspect naturel du signal de parole.

Il se peut également que la durée des phonèmes (670 ms) soit un peu longue. A titre d'exemple, Monson et al. (2011) ont utilisé une voyelle [a] isolée d'une durée de 500 ms afin d'évaluer la capacité des juges à détecter l'atténuation de l'énergie dans les HF. Ko (2021) a récemment exploré l'influence de la durée des voyelles sur l'aspect naturel de mots monosyllabiques. Cinq

expériences ont été menées auprès d'étudiants universitaires. Les résultats ont montré que les stimuli comportant une voyelle écourtée présentaient toujours un plus haut degré de naturel par rapport aux stimuli comportant une voyelle allongée. Ce constat était valable pour les paires minimales, les mots isolés et les non-mots. Concernant les non-mots, le résultat suggère que les auditeurs tendent à s'appuyer sur les règles phonotactiques[6] de leur langue pour évaluer l'aspect naturel du stimulus, plutôt que sur un exemple prototypique du mot. Par conséquent, il se peut que nos phonèmes présentent une durée plus longue que celle habituellement rencontrée et engrammée d'un point de vue phonotactique. Cela est susceptible d'avoir altéré leur aspect naturel, peu importe la modélisation utilisée. L'ajustement de la durée des phonèmes est donc un enjeu pour une éventuelle étude ultérieure fondée sur ce type de stimulus.

2.2.2. Le genre de la voix de synthèse

Nous avons souhaité investiguer l'effet du genre de la voix de synthèse sur les scores attribués par les juges à l'aspect naturel des phonèmes. Notre hypothèse a été formulée à partir des résultats de l'étude de Monson et al. (2012), qui indiquent que la partie basse des HF est plus riche dans les voix masculines tandis que la partie haute des HF est plus riche dans les voix féminines. Pour rappel, les phonèmes de la seconde tâche expérimentale étaient composés de leur spectre acoustique complet, c'est-à-dire de leurs basses et de leurs hautes fréquences. Le spectre entier des phonèmes étant transmis, nous avons supposé que les différences d'amplitude dans certaines fréquences selon le genre ne devraient pas induire de différence quant à la façon de percevoir les phonèmes. Nous avons alors émis l'hypothèse selon laquelle **les phonèmes générés avec une voix féminine et ceux générés avec une voix masculine devraient être perçus avec le même degré de naturel**. Nous n'avons pas pu démontrer de différence significative en fonction du genre de la voix de synthèse ($LR = 1.130$, $p > 0.1$). **Conformément à notre hypothèse**, les phonèmes 1D et 3D générés avec des voix féminines et des voix masculines ont été perçus avec le même degré de naturel par les juges.

2.2.3. Le type de phonème

Au regard de la distribution différente de l'énergie en HF dans les consonnes et dans les voyelles (Monson, 2011), nous avons énoncé l'hypothèse que **le score attribué pour le degré de naturel varierait selon les phonèmes**. En particulier, nous sommes partis du postulat

[6] Ensemble de règles qui spécifie les séquences de phonèmes autorisées dans une langue donnée (Zissman, 1996).

qu'une plus grande quantité d'énergie en HF pourrait accentuer la différence perçue entre les modèles acoustiques 1D et 3D. Les voyelles étant moins riches en HF, nous avons pensé que le changement de modélisation des HF (1D VS 3D) serait moins remarqué et que les voyelles 1D et 3D seraient perçues de la même manière. Nous nous sommes donc attendus à ce que **les consonnes fricatives non voisées [s], [ʃ], [f], générées avec le modèle 3D, soient considérées comme plus naturelles que les voyelles [a], [i] [u] et [ə], générées avec le modèle 3D**. Pour répondre à ces hypothèses, nous avons évalué deux types d'effet : l'effet simple du type de phonème puis l'effet d'interaction entre le réalisme du modèle acoustique et le type de phonème.

Concernant l'effet simple du type de phonème, une différence significative a pu être mise en évidence par un test du rapport de vraisemblance des modèles de liens cumulatifs (LR = 464.038, $p < 0.001$). Nous sommes en mesure de **confirmer notre première hypothèse** puisque le type de phonème exerce une influence sur le score d'aspect naturel attribué par les juges. Nous avons souhaité approfondir nos analyses en contrastant, un à un, les scores d'aspect naturel des phonèmes. Plus précisément, le score moyen de l'aspect naturel de chaque phonème a été comparé à celui des six autres phonèmes. Les phonèmes testés étaient les suivants : [a], [i], [u], [ə], [ʃ], [s], [f]. Plusieurs différences significatives entre les phonèmes ont été relevées. Presque l'ensemble des phonèmes est considéré comme significativement moins naturel que le [a], à savoir le [u], le [ə], le [f], le [s] et le [ʃ]. Seul le [i] semble être associé à un degré de naturel similaire à celui du [a].

Le [i] est perçu comme significativement plus naturel que la majorité des autres phonèmes, c'est-à-dire le [ə], le [f], le [s] et le [u]. Seuls le [a] et le [ʃ] présentent un aspect naturel équivalent à celui du [i], bien que l'aspect naturel du [ʃ] semble être légèrement moins apprécié compte-tenu la valence négative du résultat ($z = -1.116$, $p < 0.01$).

Concernant le [u], il paraît significativement moins naturel que le [a] et le [i] mais est considéré comme beaucoup plus naturel que le [f] de façon significative. Il présente un degré de naturel similaire au [ə] et au [s].

A propos du [ə], il est perçu comme significativement moins naturel que le [ʃ], le [a] et le [i], à l'exception du [f] qui semble nettement moins naturel encore. Le [s] et le [u] ont un aspect naturel similaire au [ə].

Lorsque nous considérons le [ʃ], il semble moins naturel que le [a] mais est perçu comme plus naturel que le [ə], le [s] et le [f]. Un degré de naturel équivalent a été observé avec le [i] et le [u].

Le phonème [s] est considéré comme plus naturel que le [f] par les juges mais perd en aspect naturel face au [ʃ], au [a] et au [i]. Un degré de naturel similaire est observé avec le [ə] et le [u]. Enfin, en ce qui concerne le [f], une différence significative a été relevée pour l'ensemble des contrastes réalisés. Il semblerait que la significativité s'explique en termes d'absence d'aspect naturel. Autrement dit, le [f] est considéré comme étant bien moins naturel par rapport à l'ensemble des autres phonèmes, à savoir le [ə], le [ʃ], le [a], le [i], le [s] et le [u], de façon importante.

Ces résultats montrent que, sans tenir compte du réalisme du modèle acoustique, les consonnes fricatives non voisées [s], [ʃ], [f] ne sont pas perçues comme plus naturelles que les voyelles [a], [i] [u] et [ə] par les juges. Il convient de noter que les phonèmes [f] et [s] se distinguent par un moins haut degré de naturel en comparaison aux autres phonèmes évalués.

Notre seconde hypothèse indiquait que **les consonnes fricatives non voisées [s], [ʃ], [f], générées avec le modèle 3D, seraient considérées comme plus naturelles par les juges par rapport aux voyelles [a], [i] [u] et [ə] générées avec le modèle 3D**. Afin de répondre à cette hypothèse, nous nous sommes intéressés à l'effet d'interaction entre le réalisme du modèle acoustique et le type de phonème. Investiguer cet effet d'interaction visait aussi à approfondir les résultats obtenus pour l'effet simple du phonème.

Le test du rapport de vraisemblance des modèles de liens cumulatifs n'a pas permis de mettre en évidence des différences significatives concernant l'effet d'interaction entre le modèle 3D et le type de phonème, peu importe le phonème ($LR = 6.821$, $p < 0.1$). L'association entre la modélisation 3D et un phonème particulier ne semble pas influencer son score d'aspect naturel. Nous ne sommes donc **pas en mesure d'accepter notre seconde hypothèse** puisqu'aucun phonème 3D ne paraît plus naturel que les autres d'après les juges.

Spontanément, nous aurions tendance à attribuer l'absence de différence d'aspect naturel entre les différents phonèmes au type de modélisation utilisée. Cependant, seule la modélisation 3D a été considérée dans le cadre de notre seconde hypothèse. Par ailleurs, les données montrent que cela n'est pas dû au type de modèle. Aucune différence significative n'a pu être relevée à propos de l'effet simple du réalisme du modèle acoustique ($z = 2.961$, $p < 0.1$). Les phonèmes générés par le modèle 1D et le modèle 3D ont été perçus avec un degré de naturel équivalent. La figure 22 (voir page 63) soutient cette interprétation. Elle montre des résultats plutôt aléatoires entre les deux modèles selon le phonème. Cela semble indiquer qu'il n'y a pas un modèle qui est considéré comme plus naturel que l'autre. Le seul effet simple significatif démontré concerne le phonème. Tous les phonèmes ne sont pas perçus avec le même degré de

naturel, peu importe le modèle employé. Par conséquent, il semblerait que les résultats soient davantage attribuables aux phonèmes seuls, sans prendre en compte d'autres variables. Ceci soulève la question de leur modélisation à l'aide de la synthèse articulatoire. L'usage des modèles 3D est encore peu répandu à l'heure actuelle et une marge de progression existe quant à leur développement (Arnela et al., 2019 ; Freixes et al., 2019). Ils restent simples et ne permettent d'offrir un réalisme des signaux générés qu'au niveau de la propagation des ondes acoustiques. Pour que les signaux synthétiques soient proches de signaux réels d'un point de vue naturel, il faudrait copier le processus de production de parole humaine avec une grande précision, c'est-à-dire en tenant compte des écoulements d'air, de l'ensemble des subtilités articulatoires, etc. (R. Blandin, communication personnelle, 9 août 2021). Il s'agit donc d'un travail très chronophage et nécessitant d'importants moyens computationnels.

2.2.4. La fiabilité intra-juge

De la même manière que dans la première tâche expérimentale, l'effet du moment de présentation du stimulus pendant le testing a été évalué. Il nous a permis de vérifier la fiabilité intra-juge. Chaque stimulus a été présenté deux fois, à deux moments différents du testing : celui du « test » et celui du « retest ». Nous nous attendions à ce que **les résultats des juges soient équivalents entre le test et le retest**. Les analyses statistiques n'ont pas pu montrer de différence significative à propos du moment de présentation du stimulus ($LR = 0.518$, $p > 0.1$). Les juges ont évalué de la même manière le degré de naturel des phonèmes durant la phase de test et celle du retest.

2.2.5. La fiabilité inter-juges

Notre étude étant fondée sur des expériences perceptives, évaluer la fiabilité inter-juges nous a semblé essentiel afin de vérifier le degré d'accord entre les juges. Wuyts et al. (1999) insistent d'ailleurs sur l'importance de considérer la fiabilité inter-juges lorsque des échelles perceptives sont utilisées. Un ICC égal à 0.919 a été obtenu. D'après Koo et Li (2016), ce résultat correspond à une fiabilité inter-juges excellente. Nous sommes en mesure de **confirmer** notre hypothèse, selon laquelle les résultats **entre les juges sont équivalents concernant le score de degré de naturel attribué aux phonèmes**.

3. Limites, biais méthodologiques et perspectives

Plusieurs limites et biais méthodologiques sont à mentionner dans le cadre de ce mémoire. Les résultats exposés ci-dessus et leur interprétation doivent donc être considérés avec prudence. Des perspectives sont offertes pour améliorer l'ensemble de la procédure expérimentale appliquée dans ce mémoire afin de permettre de potentielles études ultérieures sur le sujet.

3.1. Informations données aux juges

Chaque juge a reçu un formulaire d'informations explicitant le contexte et les enjeux dans lesquels s'inscrit ce mémoire. Ce formulaire précise que l'étude se fonde sur un projet lié à la synthèse de la parole. Ceci est susceptible d'avoir altéré la naïveté des juges vis-à-vis des tâches perceptives proposées. Nous ne pouvons pas exclure le fait que la connaissance de cette information ait influencé le comportement perceptif des juges. Ces derniers ont pu modifier leurs attentes perceptives, résultant en un changement des références prototypiques (Papcun et al., 1989). Cette suggestion est applicable pour la seconde tâche expérimentale, au sein de laquelle les stimuli étaient plus semblables à des signaux de parole entendus au quotidien. Les juges auraient donc évalué l'aspect naturel des stimuli en opérant des comparaisons avec des signaux de parole artificiels, réduisant ou augmentant potentiellement leur sévérité vis-à-vis du score attribué. Une possible solution serait de cacher la motivation réelle de l'étude afin qu'aucun biais ne soit introduit.

3.2. Environnement de l'étude

Des biais liés à l'environnement du testing sont à mentionner. Un chauffage fixé au mur était présent dans la cabine. Il émettait de façon aléatoire un bruit de souffle continu, qui survenait parfois pendant l'audiométrie ou les tâches perceptives. Ce bruit était probablement lié au thermostat de l'ensemble du bâtiment. Il était donc impossible de prédire sa durée ou son moment d'apparition. Des précautions ont été prises mais se sont montrées insuffisantes. Elles consistaient à éteindre le chauffage complètement en amont des testings pour éviter que le bruit ne survienne pendant ceux-ci. Il conviendrait, pour des études futures, de vérifier la source du bruit du chauffage pour qu'il ne vienne pas biaiser les mesures effectuées pendant les testings. Les mesures du niveau de bruit ambiant (à 22.8 dB SPL avec le premier ordinateur et à 18.4 dB SPL avec le second), effectuées en amont des testings, ont toutefois révélé des données acceptables.

3.3. Matériel utilisé

Afin d'évaluer les seuils auditifs des juges dans les HF, un audiomètre a été développé par Xavier Kaiser, ingénieur de recherche dans les locaux du CEDIA (bureau d'études et de recherche spécialisé en acoustique et vibrations de l'ULiège). Sa conception a rencontré quelques obstacles. Avant de débuter les testings, un bruit parasite était présent pour les deux fréquences testées, à savoir 12.5 kHz et 16 kHz. Il était perceptible au début et à la fin du stimulus. Xavier Kaiser est parvenu à supprimer ce bruit parasite à 12.5 kHz, mais pas à 16 kHz. Le calendrier imposé pour l'étude ne nous a pas permis d'attendre davantage et nous avons dû débuter les testings avec le bruit parasite à 16 kHz. Il était souvent comparé à un bruit de grincement du casque par les juges, qui n'étaient pas sûrs de devoir répondre à ce type de son. Certains ne répondaient d'ailleurs pas du tout, pensant que le bruit venait des mouvements du casque. Afin de pallier ce problème, nous avons décidé de l'expliciter en amont de l'audiométrie tonale. Cependant, la principale difficulté rencontrée était l'inconsistance du signal à 16 kHz. Parfois, le bruit parasite prenait le dessus sur le stimulus initial tandis que d'autres fois, il n'était pas présent et nous pouvions nettement percevoir le stimulus. Tous les juges n'ont donc pas été soumis au même signal. Ceci nous a amené à exclure la fréquence de 16 kHz en tant que critère d'inclusion des juges dans l'étude. En outre, bien que les niveaux de sortie de l'audiomètre aient été ajustés aux valeurs requises, l'audiomètre n'a pas été étalonné de façon normalisée comme pour l'audiomètre Madsen Itera II. Par conséquent, il serait intéressant de solutionner la présence du bruit parasite à 16 kHz et de réaliser un étalonnage de l'appareil par une entreprise pour s'assurer de sa fiabilité dans des études ultérieures.

3.4. Caractéristiques de la modalité de réponse des juges

Une amélioration peut être envisagée concernant la modalité de réponse des juges au sein de la seconde tâche expérimentale. Pour rappel, les juges étaient invités à évaluer l'aspect naturel de phonèmes à l'aide d'une échelle de Likert. Celle-ci était composée de quatre points, de 0 « pas du tout naturel » à 3 « totalement naturel ». Des analyses statistiques ont indiqué le manque de variabilité de l'échelle. Un effet significatif entre les niveaux 0 et 1 ($z = -5.959$, $p < 0.001$) et les niveaux 2 et 3 ($z = 7.896$, $p < 0.001$) a été relevé. Nous ne sommes cependant pas parvenus à montrer de différence significative entre les niveaux 1 et 2 ($z = 1.195$, $p > 0.1$). Il semblerait que les échelons 1 et 2 réfèrent au même degré de naturel pour les juges. Ces résultats reflètent les constats de Knapp (1990), qui indique que les étiquettes intermédiaires

sont susceptibles d'être considérées comme similaires compte-tenu la faible distance qui les sépare. Au contraire, les étiquettes extrêmes se distinguent plus facilement.

Une alternative à cette échelle de Likert consiste en l'utilisation d'un curseur de 0 à 100, qui offre une mesure en millimètres. Remplacer l'échelle ordinale par une échelle métrique permettrait de considérer l'aspect naturel en tant que variable métrique plutôt qu'ordinale dans les analyses statistiques. L'emploi d'une échelle visuelle analogique permet l'obtention de données plus précises puisque des analyses statistiques paramétriques peuvent être envisagées, amenant à des résultats plus précis (Bishop & Herron, 2015).

VII. CONCLUSION

Ce mémoire s'est inscrit dans un projet de développement d'une synthèse articulatoire à large bande pour mieux définir le lien entre la perception de la parole et les aspects physiques et acoustiques liés à sa production dans l'entièreté du registre fréquentiel audible (de 0.02 à 20 kHz). La synthétisation de phonèmes s'est appuyée sur deux types de modélisation physique des HF : la modélisation unidimensionnelle (1D) et la modélisation tridimensionnelle (3D). La synthèse articulatoire, combinée à l'emploi d'un modèle physique tridimensionnel (3D), semble prometteuse pour générer un signal de parole proche de la parole humaine. Bien que leur usage soit encore peu répandu, les modèles 3D permettent de représenter avec plus de réalisme acoustique la gamme des hautes fréquences (HF) situées au-delà de 5 kHz. Longtemps ignorées au profit des basses fréquences (< 5 kHz), aucune étude n'est parvenue à démontrer l'inutilité perceptive des HF. Dans ce contexte, nous avons cherché à comprendre la contribution des HF pour l'aspect naturel de la parole en fonction de modélisations physiques différentes.

Un premier objectif visait à évaluer la sensibilité auditive de jeunes adultes à divers degrés de réalisme dans la synthèse des HF. Pour ce faire, une tâche de discrimination de paires de phonèmes générés par le modèle 1D et par le modèle 3D a été administrée à 31 juges. Trois types de paires de modèles ont été proposées : les paires 1D-1D, 3D-3D et 1D-3D. Nous avons observé une différence significative concernant la perception de différences entre les modélisations physiques 1D et 3D. Bien que la capacité de discrimination était plus faible pour les paires 1D-3D, les juges sont parvenus à significativement percevoir des différences entre les deux degrés de réalisme employés. La capacité à correctement discriminer les paires de phonèmes était meilleure lorsque les phonèmes étaient générés avec une voix féminine. En outre, contrairement à ce que nous avions supposé, les paires 1D-3D les mieux discriminées étaient celles des voyelles [a], [i] et [u]. Les phonèmes [ə], [s] et [f] étaient, quant à eux, associés à une faible faculté de discrimination.

Un second objectif avait pour but d'approfondir les résultats de la première expérience. Il consistait à déterminer si une description correcte de la physique dans la synthèse des HF, telle que dans le modèle 3D, contribue à l'aspect naturel de la parole artificielle. Une tâche d'évaluation de l'aspect naturel de phonèmes a été proposée aux 31 juges. Nous ne sommes pas parvenus à mettre en évidence un effet significatif concernant le modèle utilisé et le genre de la voix de synthèse. Les phonèmes 1D et 3D, générés avec une voix féminine et une voix masculine, ont été considérés avec un degré de naturel similaire. De plus, un effet significatif

du phonème a été mis en évidence, qui montre que le score d'aspect naturel varie selon le phonème évalué. Il semblerait que l'effet du phonème soit plus important que l'effet du modèle. Ceci soulève donc la question de la modélisation 3D des phonèmes à l'aide de la synthèse articulatoire, qui permet à ce jour d'offrir des signaux réalistes uniquement du point de vue acoustique. Les modèles 3D nécessitent encore d'être améliorés afin de réduire la distance entre le réalisme acoustique des signaux de parole artificiels et le réalisme perceptif des signaux de parole humains.

Alors que certaines études ont déjà comparé les différences entre les modèles 1D et 3D d'un point de vue acoustique (Arnela et al., 2019 ; Freixes et al., 2019), aucune étude ne semblait avoir investigué l'effet de ces différences acoustiques sur la perception de la parole auprès d'auditeurs réels. Notre étude a finalement permis de mettre en lumière la sensibilité auditive de jeunes adultes à une modélisation 3D des HF, plus complexe et réaliste acoustiquement, mais n'est pas parvenue à définir l'apport du modèle 3D pour l'aspect naturel des phonèmes.

L'ensemble de nos résultats est à considérer avec précaution au regard des divers biais méthodologiques relevés pour cette étude. Afin d'améliorer la recherche menée, nous recommandons aux études ultérieures :

- D'utiliser un audiomètre étalonné dans les HF pour assurer la fiabilité des résultats des seuils auditifs des juges et, par conséquent, des résultats aux tâches perceptives.
- D'employer de la parole connectée, tels que des mots monosyllabiques plutôt que des phonèmes isolés, pour mettre en évidence des différences entre les modèles 1D et 3D à propos de l'aspect naturel des signaux de parole.
- D'ajuster la durée des phonèmes pour que celle-ci n'interfère pas dans l'évaluation de l'aspect naturel avec le type de modélisation employée (1D ou 3D), dans le cas où des phonèmes isolés sont utilisés en tant que stimuli.
- D'utiliser une échelle d'évaluation de l'aspect naturel des stimuli avec un curseur de 0 à 100 pour gagner en fiabilité et précision lors des analyses statistiques.

Cette étude reste exploratoire dans l'investigation du rôle des HF pour la perception de la parole selon des modélisations physiques différentes. Des études supplémentaires sont donc nécessaires pour améliorer notre recherche et définir plus précisément les enjeux d'une modélisation 3D des HF pour la perception de la parole.

VIII. BIBLIOGRAPHIE

Anand, S., & Stepp, C. E. (2015). Listener perception of monopitch, naturalness, and intelligibility for speakers with Parkinson's disease. *Journal of Speech, Language, and Hearing Research*, *58*(4), 1134-1144. https://doi.org/10.1044/2015_JSLHR-S-14-0243

Apoux, F., & Bacon, S. P. (2004). Relative importance of temporal information in various frequency regions for consonant identification in quiet and in noise. *The Journal of the Acoustical Society of America*, *116*(3), 1671-1680. https://doi.org/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), 2173-2182. https://doi.org/10.1109/TASLP.2019.2942439

Bachorowski, J. A., & Owren, M. J. (1999). Acoustic correlates of talker sex and individual talker identity are present in a short vowel segment produced in running speech. *The Journal of the Acoustical Society of America*, *106*(2), 1054-1063. https://doi.org/10.1121/1.427115

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. *The Journal of the Acoustical Society of America*, *129*(2), 852-863. https://doi.org/10.1121/1.3523476

Baird, A., Parada-Cabaleiro, E., Hantke, S., Burkhardt, F., Cummins, N., & Schuller, B. (2018, September 2-6). *The perception and analysis of the likeability and human likeness of synthesized speech* [Paper presentation]. Interspeech, Hyderabad, Pakistan. https://doi.org/10.21437/Interspeech.2018-1093

Bauer, D., Birkholz, P., Kannampuzha, J., & Kröger, B. J. (2010). Evaluation of articulatory speech synthesis: a perception study. *36th Deutsche Jahrestagung fr Akustik* 1003-1004.

Baken, R. J., & Orlikoff, R. F. (2000). *Clinical measurement of speech and voice* (2nd ed.). Singular Thomson Learning.

Baumann, O., & Belin, P. (2010). Perceptual scaling of voice identity: Common dimensions for different vowels and speakers. *Psychology Research*, 74, 110– 120. https://doi.org/10.1007/s00426-008-0185-z

Belin, P., Fecteau, S., & Bedard, C. (2004). Thinking the voice: Neural correlates of voice perception. *Trends in cognitive sciences*, *8*(3), 129-135. https://doi.org/10.1016/j.tics.2004.01.008

Belin, P., & Grosbras, M. H. (2010). Before speech: Cerebral voice processing in infants. *Neuron*, *65*(6), 733-735. https://doi.org/10.1016/j.neuron.2010.03.018

Belin, P., Bestelmeyer, P. E., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology*, *102*(4), 711-725. https://doi.org/10.1111/j.2044-8295.2011.02041.x

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), 353-363. https://doi.org/10.1121/1.1926107

Birkholz, P., Jackèl, D., and Kröger, B. J. (2006). Construction and control of a three-dimensional vocal tract model. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, 873–876, Toulouse, France.

Birkholz, P. (2013). Modeling consonant-vowel coarticulation for articulatory speech synthesis. *PLOS ONE*, *8*(4), 1-17. https://doi.org/10.1371/journal.pone.0060603

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*(3), 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*, 116-127. https://doi.org/10.1016/j.csl.2016.06.004

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

Bishop, P. A., & Herron, R. L. (2015). Use and misuse of the Likert item responses and other ordinal measures. *International journal of exercise science*, *8*(3), 297-302.

Blandin, R., Arnela, M., Laboissière, R., Pelorson, X., Guasch, O., Hirtum, A. V., & Laval, X. (2015). Effects of higher order propagation modes in vocal tract like geometries. *The Journal of the Acoustical Society of America*, *137*(2), 832-843. https://doi.org/10.1121/1.4906166

Boatman, D., Lesser, R. P., & Gordon, B. (1995). Auditory speech processing in the left temporal lobe: An electrical interference study. *Brain and language*, *51*(2), 269-290. https://doi.org/10.1006/brln.1995.1061

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.

Bureau International d'Audiophonologie. (n.d.). Recommandation biap 02/1 bis : Classification audiométrique des déficiences auditives. *BIAP – Bureau International d'Audiophonologie*. Consulté le 9 avril 2020, sur https://www.biap.org/en/component/content/article/65-recommendations/ct-2-classification/5-biap-recommendation-021-bis

Castellanos, A., Benedí, J. M., & Casacuberta, F. (1996). An analysis of general acoustic-phonetic features for Spanish speech produced with the Lombard effect. *Speech Communication*, *20*(1-2), 23-35. https://doi.org/10.1016/S0167-6393(96)00042-8

Castellengo, M. (2015). *Ecoute musicale et acoustique : Avec 420 sons et leurs audiogrammes décryptés*. Eyrolles.

Collet, L. (1994). Les oto-émissions acoustiques chez l'humain. *Archives Internationales de Physiologie, de Biochimie et de Biophysique*, *102*(4), 45-53. https://doi.org/10.3109/13813459109045392

Delvaux, V., & Pillot-Loiseau, C. (2019). Perceptual judgment of voice quality in nondysphonic french speakers: Effect of task-, speaker- and listener-related variables. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2019.02.013

Diehl, R. L., Lotto, A. J., & Holt, L. L. (2004). Speech perception. *Annual Review of Psychology*, 55, 149-179. https://doi.org/10.1146/annurev.psych.55.090902.142028

Dutoit, T. (1997). High-quality text-to-speech synthesis: An overview. *Journal Of Electrical And Electronics Engineering Australia*, *17*, 25-36.

Dutoit, T., Couvreur, L., Malfrere, F., Pagel, V., & Ris, C. (2002). Synthèse vocale et reconnaisance de la parole : Droites gauches et mondes paralleles. In *Actes du 6e Congrès Francais d'Acoustique*, Lille, France.

Englert, S. E. (1986). *The speech spectrum and its relationship to intelligibility of speech* [Doctoral dissertation, University of Wyoming]. ProQuest. https://www.proquest.com/openview/54ffffb82b06ca69761c6a9423473141/1?cbl=18750&diss=y&pq-origsite=gscholar

Fitch, W. T. (1997). Vocal tract length and formant frequency dispersion correlate with body size in rhesus macaques. *The Journal of the Acoustical Society of America*, *102*(2), 1213-1222. https://doi.org/10.1121/1.421048

Freixes, M., Arnela, M., Socoró, J. C., Pujol, F. A., & Guasch, O. (2018, November 21-23). *Influence of tense, modal and lax phonation on the three-dimensional finite element synthesis of vowel /a/* [Paper presentation]. IberSPEECH 2018, Barcelona, Spain.

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

Füllgrabe, C., Baer, T., Stone, M. A., & Moore, B. C. (2010). Preliminary evaluation of a method for fitting hearing aids with extended bandwidth. *International Journal of Audiology*, *49*(10), 741-753. https://doi.org/10.3109/14992027.2010.495084

Garnier, M., & Henrich, N. (2014). Speaking in noise: How does the Lombard effect improve acoustic contrasts between speech and ambient noise?. *Computer Speech & Language*, *28*(2), 580-597. https://doi.org/10.1016/j.csl.2013.07.005

Garnier, M., Ménard, L., & Alexandre, B. (2018). Hyper-articulation in Lombard speech: An active communicative strategy to enhance visible speech cues?. *The Journal of the Acoustical Society of America*, *144*(2), 1059-1074. https://doi.org/10.1121/1.5051321

Gordon-Salant, S. (2005). Hearing loss and aging: New research findings and clinical implications. *Journal of Rehabilitation Research & Development*, *42*(4), 9-24. https://doi.org/10.1682/jrrd.2005.01.0006

Hickok, G., & Poeppel, D. (2000). Towards a functional neuroanatomy of speech perception. *Trends in cognitive sciences*, *4*(4), 131-138. https://doi.org/10.1016/S1364-6613(00)01463-7

Hillenbrand, J. (1987). A methodological study of perturbation and additive noise in synthetically generated voice signals. *Journal of Speech, Language, and Hearing Research*, *30*(4), 448-461.

Holmes, J., & Holmes, W. (2001). *Speech Synthesis and Recognition* (2nd ed.). Taylor & Francis.

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

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*, 107922. https://doi.org/10.1016/j.heares.2020.107922

Johnson, K. (2003). *Acoustic & auditory phonetics* (2nd ed.). Blackwell Publishing.

Kacha, A., Grenez, F., & Schoentgen, J. (2005, September 4-8). *Voice quality assessment by means of comparative judgments of speech tokens* [Paper presentation]. Interspeech, Lisbon, Portugal.

Khan, R. A., & Chitode, J. S. (2016). Concatenative speech synthesis: A Review. *International Journal of Computer Applications*, *136*(3), 1-6. https://doi.org/10.5120/ijca2016907992

Knapp, T. R. (1990). Treating ordinal scales as interval scales: An attempt to resolve the controversy. *Nursing research*, *39*(2), 121-123.

Ko, E. S. (2021). Effects of vowel duration on the perceived naturalness of English monosyllabic words ending in a stop: Some preliminary findings. *Phonetics and Speech Sciences, 13*(2), 37-44. https://doi.org/10.13064/KSSS.2021.13.2.037

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, *15*(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Kreiman, J., Gerratt, B. R., & Precoda, K. (1990). Listener experience and perception of voice quality. *Journal of Speech, Language, and Hearing Research*, *33*(1), 103-115. https://doi.org/10.1044/jshr.3301.103

Kreiman, J., & Papcun, G. (1991). Comparing discrimination and recognition of unfamiliar voices. *Speech Communication*, *10*(3), 265-275. https://doi.org/10.1016/0167-6393(91)90016-M

Kreiman, J., Gerratt, B. R., & Berke, G. S. (1994). The multidimensional nature of pathologic vocal quality. *The Journal of the Acoustical Society of America*, *96*(3), 1291-1302. https://doi.org/10.1121/1.410277

Kreiman, J., & Gerratt, B. R. (1996). The perceptual structure of pathologic voice quality. *The Journal of the Acoustical Society of America*, *100*(3), 1787-1795. https://doi.org/10.1121/1.416074

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. Wiley-Blackwell. https://doi.org/10.1002/9781444395068

Kröger, B. (1992). Minimal rules for articulatory speech synthesis. *Proceedings of EUSIPCO92 (1)*, 331-334.

Kuligowska, K., Kisielewicz, P., & Włodarz, A. (2018). Speech synthesis systems: Disadvantages and limitations. *International Journal of Engineering & Technology*, *7* (2.28), 234-239. https://doi.org/10.14419/ijet.v7i2.28.12933

Larrouy-Maestri, P., Wang, X., Nunes, R. V., & Poeppel, D. (in press). Are you your own best judge? On the self-evaluation of singing. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2021.03.028

Latinus, M., & Belin, P. (2011). Human voice perception. *Current Biology*, *21*(4), 143-145. https://doi.org/10.1016/j.cub.2010.12.033

Latinus, M., McAleer, P., Bestelmeyer, P. E., & Belin, P. (2013). Norm-based coding of voice identity in human auditory cortex. *Current Biology*, *23*(12), 1075-1080. https://doi.org/10.1016/j.cub.2013.04.055

Lippmann, R. P. (1996). Accurate consonant perception without mid-frequency speech energy. *IEEE Transactions on Speech and Audio Processing*, *4*(1), 66. https://doi.org/10.1109/TSA.1996.481454

Lukose, S., & Upadhya, S. S. (2017). Text to speech synthesizer-formant synthesis. *2017 International Conference on Nascent Technologies in Engineering (ICNTE-2017)*, 1-4. https://doi.org/10.1109/ICNTE.2017.7947945

Margolis, R. H., Saly, G. L., & Hunter, L. L. (2000). High-frequency hearing loss and wideband middle ear impedance in children with otitis media histories. *Ear and hearing*, *21*(3), 206-211. https://doi.org/10.1097/00003446-200006000-00003

Monson, B. B. (2011). *High-Frequency Energy in Singing and Speech* [Doctoral dissertation, University of Arizona]. The University of Arizona. https://repository.arizona.edu/handle/10150/202695

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), 2263-2268. https://doi.org/10.1121/1.3557033

Monson, B. B., Lotto, A. J., & Story, B. H. (2012). Analysis of high-frequency energy in long-term average spectra of singing, speech, and voiceless fricatives. *The Journal of the Acoustical Society of America*, *132*(3), 1754-1764. https://doi.org/10.1121/1.4742724

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), 1-11. https://doi.org/10.3389/fpsyg.2014.00587

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), 1-7. https://doi.org/10.3389/fpsyg.2014.01239

Monson, B., & Buss, E. (2019). Does extended high-frequency hearing matter in real-world listening?. *The Hearing Journal*, *72*(12), 30-32. https://doi.org/10.1097/01.HJ.0000616140.40371.b9

Monson, B. B., & Caravello, J. (2019). The maximum audible low-pass cutoff frequency for speech. *The Journal of the Acoustical Society of America*, *146*(6), 496-501. https://doi.org/10.1121/1.5140032

Monson, B. B., Rock, J., Schulz, A., Hoffman, E., & Buss, E. (2019). Ecological cocktail party listening reveals the utility of extended high-frequency hearing. *Hearing research*, *381*. https://doi.org/10.1016/j.heares.2019.107773

Moore, B. C., & Tan, C. T. (2003). Perceived naturalness of spectrally distorted speech and music. *The Journal of the Acoustical Society of America*, *114*(1), 408-419. https://doi: 10.1121/1.1577552

Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar voices. *The Journal of the Acoustical Society of America*, *85*(2), 913-925. https://doi.org/10.1121/1.397564

Remacle, A., Schoentgen, J., Finck, C., Bodson, A., & Morsomme, D. (2014). Impact of vocal load on breathiness: Perceptual evaluation. *Logopedics Phoniatrics Vocology*, *39*(3), 139-146. https://doi.org/10.3109/14015439.2014.884161

Robertson, S. J., & Thomson, F. (1999). *Rééduquer les dysarthriques*. Ortho Edition.

Rodríguez Valiente, A., Trinidad, A., García Berrocal, J. R., Górriz, C., & Ramirez Camacho, R. (2014). Extended high-frequency (9–20 kHz) audiometry reference thresholds in 645 healthy subjects. *International journal of audiology*, *53*(8), 531-545. https://doi.org/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), 837-846.

Scully, C. (1990). Articulatory Synthesis. *Speech Production and Speech Modelling, 55*, 151-186.

Shadle, C. H., & Scully, C. (1995). An articulatory-acoustic-aerodynamic analysisof [s] in VCV sequences. *Journal of phonetics*, *23*(1-2), 53-66. https://doi.org/10.1016/S0095-4470(95)80032-8

Shadle, C. H., & Damper, R. I. (2001, August 29-September 1). *Prospects for Articulatory Synthesis: A Position Paper* [Paper presentation]. Proceedings of 4th ISCA Workshop on Speech Synthesis, Pitlochry, Scotland.

Shroeter, J. (2005). Voice Modification for Applications in Speech Synthesis. *AT&T Labs – Research*, 1-20.

Sicard, E., & Menin-Sicard, A. (2021). Le spectrogramme et son application en clinique orthophonique. *Archive ouverte HAL*, 1-26.

Sondhi, M., & Schroeter, J. (1987). A hybrid time-frequency domain articulatory speech synthesizer. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *35*(7), 955-967. https://doi.org/10.1109/TASSP.1987.1165240

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. *The Journal of the Acoustical Society of America*, *110*(4), 2183-2190. https://doi.org/10.1121/1.1400757

Stylianou, Y. (2001). Applying the harmonic plus noise model in concatenative speech synthesis. *IEEE Transactions on speech and audio processing*, *9*(1), 21-29. https://doi.org/10.1109/89.890068

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)*, 67-70. https://doi.org/10.1109/WOSSPA.2011.5931414

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), 1-10. https://doi.org/10.1007/s00221-004-2044-5

Teston, B. (2004). L'évaluation instrumentale des dysphonies: État actuel et perspectives. In A. Giovanni (Ed.), *Le bilan d'une dysphonie: État actuel et perspectives* (pp. 105-169). Solal.

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). Elsevier.

Vaissière, J. (2011). Les organes de la parole. In J. Vaissière (Ed.), *La phonétique* (pp. 45-56). Presses Universitaires de France.

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), 65-70. https://doi.org/10.1121/1.4903917

Wuyts, F. L., De Bodt, M. S., & Van de Heyning, P. H. (1999). Is the reliability of a visual analog scale higher than an ordinal scale? An experiment with the GRBAS scale for the perceptual evaluation of dysphonia. *Journal of Voice*, *13*(4), 508-517. https://doi.org/10.1016/S0892-1997(99)80006-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), 1-7. https://doi.org/10.1073/pnas.1903315116

IX. ANNEXES

Annexe 1 : Recherche documentaire

Afin de recenser les sources scientifiques mentionnées dans ce mémoire, des bases de données, des catalogues bibliothécaires et des moteurs de recherche ont été consultés. Dans un premier temps, la question de recherche a été transformée en question PICO en vue de rechercher des données probantes et adaptées. La question PICO est la suivante : « Chez des adultes ayant une audition normale (P), est-ce que la parole de synthèse ayant deux niveaux de réalisme en hautes fréquences (I) est perçue différemment (O) de la parole naturelle (C) ? ».
La recherche documentaire s'est fondée sur l'introduction de mots-clés abordant les domaines de l'analyse perceptive de la parole, des méthodes de synthèse de la parole et des hautes fréquences :

- Analyse perceptive de la parole : *speech perception ; voice ; voice perception ; prototypical voice ; prototype ; acoustic features*
- Méthodes de synthèse de la parole : *speech synthesis ; text-to-speech ; formant synthesis ; concatenative synthesis ; articulatory synthesis ; speech naturalness ; speech intelligibility ; fricatives ; consonants ; vowels*
- Hautes fréquences : *high frequencies ; extended high frequencies ; naturalness ; intelligibility ; human auditory sensitivity ; age ; gender*

Les bases de données, catalogues bibliothécaires et moteurs de recherche cités ci-dessous sont les principales qui ont été utilisées pour l'élaboration de ce travail. Pour chaque mot-clé, différents opérateurs booléens ont été utilisés.

- Medline
- Uliège library
- ProQuest
- Pubmed
- Google scholar

En outre, certaines sources ont été recueillies par la consultation des bibliographies des articles et des citations d'autres sources dans les articles. Des livres ont également été empruntées à la bibliothèque universitaire de l'ULiège en vue d'enrichir les références à partir desquelles ce mémoire a pu être mené.

Annexe 2 : Questionnaire anamnestique

QUESTIONNAIRE ANAMNESTIQUE

Initiales (prénom, nom) :

Genre : ☐ Homme ☐ Femme ☐ Autre

Age au moment du test :

DONNÉES SCOLAIRES :

- En quelle année d'études êtes-vous inscrit actuellement ? _____
- Dans quelle filière êtes-vous inscrit à l'Université de Liège ?

  ☐ Logopédie ☐ Psychologie

DONNÉES LANGAGIÈRES :

- Quelle est votre langue maternelle ?

  _____

  _____

- Etes-vous bilingue ? ☐ Oui ☐ Non

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

  _____

  _____

- Avez-vous appris d'autres langues ? ☐ Oui ☐ Non

  Si OUI, pouvez-vous indiquer lesquelles ?

  _____

  _____

DONNÉES AUDITIVES :

- 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 ?

  _____

  _____

- Avez-vous déjà souffert d'acouphènes ? ☐ Oui ☐ Non

  Si OUI, à quel âge ?

  _____

  _____

  Si OUI, dans quelles circonstances ces acouphènes sont-ils apparus (par exemple, à la suite d'un concert) ?

  _____

  _____

  Si OUI, la durée de la présence de ces acouphènes est-elle temporaire ou permanente ?

  ☐ Temporaire ☐ Permanente

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

DONNÉES NEUROPSYCHOLOGIQUES :

- Rencontrez-vous des difficultés pour maintenir votre attention et votre concentration lors de la réalisation d'une tâche ? ☐ Oui ☐ Non

- Un diagnostic de TDA/H (trouble du déficit de l'attention avec ou sans hyperactivité) a-t-il été posé par un professionnel de la santé ? ☐ Oui ☐ Non

DONNÉES CONCERNANT L'EXPERTISE MUSICALE :

- Avez-vous déjà fait de la musique/du chant ? ☐ Oui ☐ Non

  Si OUI, quelle est votre formation musicale extra-scolaire ? (cochez la (les) case(s) adéquate(s))

  ☐ Conservatoire (niveau supérieur) ☐ Académie de musique (niveau secondaire) ☐ Autre
  ☐ Cours particuliers ☐ Autodidacte

  Quel(s) instrument(s) avez-vous appris ?

Instrument 1 :	Nombre d'années d'apprentissage formel :	Âge de début :
Instrument 2 :	Nombre d'années d'apprentissage formel :	Âge de début :
Instrument 3 :	Nombre d'années d'apprentissage formel :	Âge de début :
Solfège : ☐ Oui ☐ Non	Nombre d'années d'apprentissage formel :	Âge de début :

Chant lyrique : ☐ Oui ☐ Non Nombre d'années d'apprentissage formel : Âge de début :
Chant en chorale : ☐ Oui ☐ Non Nombre d'années d'apprentissage formel : Âge de début :
Chant de variété : ☐ Oui ☐ Non Nombre d'années d'apprentissage formel : Âge de début :

Pouvez-vous indiquer le nombre d'années durant lesquelles vous pratiquez/avez pratiqué la musique (instrument/chant) ?

_____

_____

Si vous avez arrêté de pratiquer la musique (instrument/chant), pouvez-vous indiquer depuis combien de temps (en mois ou en années) ?

_____

_____

Si vous pratiquez toujours la musique (instrument/chant), combien d'heures pratiquez-vous par semaine en moyenne ?

_____

_____

- Pour tous *(que vous ayez fait de la musique ou pas)*

Quel est votre temps d'écoute active de musique (tous styles musicaux confondus) (par exemple, mettre un disque) ?

☐ Moins d'une heure par jour
☐ Entre 1 et 2 heures par jour
☐ Plus de 2 heures par jour

Combien de jours par semaine ? _____

Quel est votre temps d'écoute passive de musique (par exemples, via la télévision, la radio, un fond musical...) ?

☐ Moins d'une heure par jour
☐ Entre 1 et 2 heures par jour
☐ Plus de 2 heures par jour

Combien de jours par semaine ? _____

Quel(s) style(s) de musique écoutez-vous ? (plusieurs réponses peuvent être sélectionnées)

☐ Musique vocale ☐ Musique instrumentale

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 3 : Rapport d'étalonnage de l'audiomètre Madsen Itera II

sonova
HEAR THE WORLD

Client Nr.

Bon Nr. Audiologie: REPB 190170

Client Name: CHU Liège

Address: Av. de l'Hopital, 1 Liège

Date: 07/08/2019

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

Order | Customer Nr.

Brand: Madsen Otometrics | Type: ITERA | Class | SN: 202163

Description: Vérification générale étalonnage

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

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 ☐	75,00 EUR
Transport	
Total (excl.TVA)	75,00 EUR

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	On Site ☒ Labo ☒	Warranty ☐ Contract ☐

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

Annexe 4 : Seuils auditifs obtenus avec l'audiométrie tonale pour chaque juge

Juge	Oreille droite							Oreille gauche						
	0.5 kHz	1 kHz	2 kHz	4 kHz	8 kHz	12.5 kHz	16 kHz	0.5 kHz	1 kHz	2 kHz	4 kHz	8 kHz	12.5 kHz	16 kHz
1	10	0	0	0	5	0	25	0	0	0	0	10	0	25
2	0	0	0	0	5	0	50	0	0	0	0	0	0	50
3	0	0	0	0	5	0	25	0	0	0	0	0	10	10
4	0	0	0	0	0	0	5	0	0	0	0	0	0	5
5	10	0	5	0	0	0	15	5	0	0	0	5	0	20
6	5	0	0	0	10	0	25	10	0	0	0	10	0	25
7	5	0	0	0	0	5	30	0	0	0	0	0	20	30
8	5	0	0	0	5	10	10	0	0	0	0	0	0	20
9	5	0	0	0	5	0	10	0	0	0	0	15	0	10
10	10	0	0	0	0	0	15	10	0	0	0	0	0	30
11	0	5	0	0	5	10	35	0	0	0	0	5	0	35
12	0	0	0	0	0	0	5	0	0	0	0	0	0	5
13	5	0	0	0	0	5	35	5	0	0	0	0	0	35
14	5	0	0	0	5	25	35	0	0	0	0	10	10	35
15	5	0	0	0	0	0	15	0	0	0	0	0	0	10
16	5	0	0	0	10	0	20	0	0	0	0	5	0	10
17	0	0	0	0	15	15	35	0	0	0	0	5	5	30
18	5	5	5	0	5	0	10	0	5	0	0	0	0	30
19	5	0	0	0	0	0	10	0	0	0	0	0	0	20
20	0	0	0	0	0	0	10	0	0	0	0	5	0	15
21	0	0	0	0	0	0	30	0	0	0	0	5	0	40
22	0	0	0	0	5	15	30	0	0	0	0	0	10	30
23	0	5	0	0	0	0	35	0	0	0	0	0	5	25
24	5	0	0	0	0	0	30	5	0	0	0	0	0	30
25	0	0	0	0	5	0	20	0	0	0	0	5	0	15
26	10	0	10	0	0	10	35	10	0	10	0	0	0	35
27	10	0	0	0	0	0	30	5	0	0	0	0	10	35
28	0	0	0	0	0	0	30	0	0	0	0	0	0	30

29	5	0	0	0	0	0	35	0	0	0	0	0	0	30
30	5	5	0	0	0	0	10	5	5	0	0	0	0	15
31	0	0	0	0	10	0	25	0	0	0	0	15	0	30

Annexe 5 : Protocole du testing

PROTOCOLE DU TESTING

1) **Aller chercher le juge dans le hall**

2) **L'inviter à se désinfecter les mains à l'aide du gel hydroalcoolique présent dans le hall**

3) **Lui donner les papiers**
   - Formulaire de consentement : préciser qu'il est libre de partir à tout moment s'il ne souhaite plus participer
   - Formulaire d'informations aux volontaires

4) **Mettre les téléphones et l'ordinateur en mode avion**

5) **Complétion du questionnaire anamnestique**

6) **Audiométrie tonale**

   1. Correctement installer le juge :
      - Confortable ?
      - Retirer les lunettes ; aides auditives ; boucles d'oreilles si besoin
      - ∅ bonbon/chewing-gum
      - Regarder droit devant soi + se concentrer

   2. Donner la consigne :
      « Vous allez écouter des sons dans le casque. Lorsque vous entendez, levez la main bien haut sinon je ne la verrai pas. Si vous n'entendez pas, ne faites rien. Attention que vous devez répondre pour un son et pas une vibration. Nous allons faire une oreille puis l'autre. »

      « Avant de commencer, nous allons faire un essai. Est-ce que vous entendez ? Dans quelle oreille entendez-vous le son ? »

      → Jouer le son d'essai à 40 dB à 1000 Hz à gauche puis à droite

   3. Correctement positionner le casque, en regard du conduit auditif externe (**!** Pas de cheveux entre l'oreille et le casque **!**) → rouge = droit // gauche = bleu

   4. Placer le cache (valise de l'audiomètre) entre le participant et l'audiomètre (**!** placer mes deux mains sur l'audiomètre pour que seuls mes doigts bougent **!**)

   5. Commencer à tester à 25 dB à 1000 Hz
      → Descendre par pas de 5 dB jusqu'à ce que le participant n'entende plus
      → Remonter de 5 dB pour vérifier que le participant entende bien le son au seuil indiqué
      → Si entente, redescendre de 5 dB pour s'assurer que le participant n'entend plus

      Ordre des fréquences testées : 1000 ; 2000 ; 4000 ; 8000 ; 12500 ; 16000 ; 500 Hz
      → 1000 ; 2000 ; 4000 ; 8000 ; 500 Hz avec l'audiomètre MADSEN
      → 12 500 ; 16 000 Hz avec l'audiomètre du CEDIA

   6. Retirer l'audiomètre de la table et le poser sur le côté

7) **Pause de 5 minutes à l'extérieur de la cabine audiométrique (si nécessaire pour le participant)**

8) <u>Tâche 1 (sons différents – identiques)</u>

1. Brancher le haut-parleur à l'ordinateur
2. Mettre le son de l'ordinateur au maximum
3. Ouvrir le logiciel Octave
4. Lancer la tâche « pair comparison »
5. Encoder le numéro du juge
6. Donner la consigne :

« Vous allez écouter 2 sons et vous devrez indiquer si vous avez l'impression qu'ils sont identiques ou différents. Voici à quoi ressemble 1 paire de sons identiques (la jouer) et voici à quoi ressemble 1 paire de sons différents (la jouer). Il peut s'agir de tout type de différence. Pour écouter les sons, vous devez cliquer sur le bouton « son 1 » et le bouton « son 2 ». Pour passer à la paire suivante, vous devez cliquer sur « paire suivante ». Les sons pourront vous paraître particuliers car ce sont des sons que nous n'avons pas l'habitude d'entendre dans notre quotidien. Vous pouvez également les écouter autant de fois que vous le souhaitez mais lorsque vous aurez cliqué sur « paire suivante », il ne sera plus possible de revenir en arrière. Nous allons d'abord faire un entraînement avec 2 paires de sons, puis vous aurez 84 paires de sons à juger. Si vous avez des questions à la fin de cet entraînement, n'hésitez pas à les poser. ».

7. Enregistrer le fichier CVS sous format EXCEL

9) <u>Tâche 2 (aspect naturel)</u>

1. Vérifier que le son de l'ordinateur est toujours au maximum
2. Lancer la tâche « eva interface 4scale »
3. Encoder le numéro du juge
4. Donner la consigne :

« Vous allez écouter des voyelles et des consonnes et vous devrez évaluer leur aspect naturel sur une échelle de 0 à 3, 0 signifiant « pas du tout naturel » et 3 signifiant « totalement naturel ». Par aspect naturel, nous faisons référence à une voix que nous pouvons entendre dans la vie de tous les jours. Il n'y a pas de bonne ou de mauvaise réponse, c'est selon votre ressenti. Pour écouter le son, vous devez cliquer sur le bouton « son ». Pour passer au son suivant, vous devez cliquer sur « suivant ». Vous pouvez écouter le son autant de fois que vous le souhaitez mais lorsque vous aurez cliqué sur « suivant », il ne sera plus possible de revenir en arrière. Nous allons à nouveau d'abord faire un entraînement avec 2 sons, puis vous aurez 56 sons à évaluer. Si vous avez des questions à la fin de cet entraînement, n'hésitez pas à les poser. ».

5. Enregistrer le fichier CVS sous format EXCEL

10) <u>Accompagner le juge dans le hall du bâtiment</u>

11) <u>Désinfection du local</u>

1. Laisser les portes de la cabine audiométrique ouvertes

2. Désinfecter :
    - Table
    - Chaise
    - Souris
    - Clavier de l'ordinateur
    - Audiomètres
    - Casques des audiomètres
    - Poignées de porte
    - Stylo utilisé pour compléter le questionnaire anamnestique

→ Espacer les rendez-vous de 20 minutes entre les juges

Résumé

Introduction et objectifs : Les hautes fréquences (HF) de la parole (> 5 kHz) ont été ignorées dans la majorité des recherches jusque dans les années 2010, au profit de l'étude de l'énergie en basse fréquence (< 5 kHz) considérée comme suffisante pour l'intelligibilité de la parole (Boyd-Pratt & Donai, 2020 ; Monson & Caravello, 2019 ; Vitela et al., 2015). Jusqu'alors, aucune étude n'était pourtant parvenue à démontrer l'inutilité perceptive des HF (Monson, Hunter et al., 2014). La recherche de Birkholz et Drechsel (2021) a d'ailleurs suggéré leur potentiel rôle pour produire un signal de parole plus naturel. Ce mémoire s'inscrit dans un projet de développement d'une synthèse articulatoire à large bande à partir de deux types de modélisation physique des hautes fréquences (HF) : la modélisation unidimensionnelle (1D) et la modélisation tridimensionnelle (3D). Il vise à mieux comprendre et définir le lien entre la perception de la parole et les aspects physiques et acoustiques liés à sa production dans l'entièreté du registre fréquentiel audible (0.02 à 20 kHz). Une sensibilité auditive chez de jeunes adultes entre les modèles 1D et 3D pour la synthèse des HF devrait être objectivée. En outre, les stimuli générés avec le modèle 3D devraient être considérés comme plus naturels que ceux générés avec le modèle 1D, compte-tenu sa description plus complète des effets de la géométrie tridimensionnelle du tractus vocal sur ses propriétés acoustiques (Arnela et al., 2019).
Méthodologie : Après avoir complété un questionnaire anamnestique et réalisé une audiométrie tonale, 31 juges ont réalisé deux tâches perceptives. Une première tâche de discrimination de paires de stimuli a été proposée, au sein de laquelle les juges devaient indiquer si la paire était identique ou différente. La seconde tâche expérimentale consistait à évaluer l'aspect naturel de phonèmes sur une échelle de Likert allant de 0 « pas du tout naturel » à 3 « totalement naturel ». Ces expériences nous ont permis de répondre à plusieurs hypothèses concernant la modélisation utilisée, le genre de la voix de synthèse, le type de phonème et les fiabilités intra- et inter-juges.
Résultats et conclusions : Un effet significatif concernant la perception de différences entre les modèles physiques 1D et 3D, avec une capacité de discrimination plus faible pour les paires 1D-3D, a été relevé au sein de la première tâche expérimentale. Aucun effet significatif concernant le modèle utilisé n'a pu être montré pour la seconde tâche, au sein de laquelle les stimuli 1D et 3D ont été considérés avec un degré de naturel similaire. En outre, peu importe la modélisation employée, nous avons constaté que l'aspect naturel dépend du phonème. Cette étude reste exploratoire dans la définition du rôle des HF pour la perception de la parole selon des modélisations physiques différentes. Des études confirmatoires sont donc nécessaires.