

Université Sorbonne Nouvelle — Paris 3
Institut de Linguistique et Phonétique Générales et Appliquées (ILPGA)
Laboratoire de Phonétique et Phonologie, UMR 7018

Variation du type de phonation et sa perception selon le locuteur

MEMOIRE DE MASTER 1/2 RECHERCHE MENTION SCIENCES DU LANGAGE
Parcours : Phonétique et Phonologie

Présenté par
Carole MILLOT
21800813

Devant le jury composé de
Nicolas Audibert (MCF)
Cédric Gendrot (MCF)

Anne Hermes (Chargée de Recherche au CNRS)

Sous la direction de M.Cédric GENDROT

Année universitaire 2021-2022

Déclaration sur l'honneur

Je, soussignée, Carole MILLOT, déclare avoir rédigé ce travail sans aides extérieures ni sources autres que celles qui sont citées. Toutes les utilisations de textes préexistants, publiés ou non, y compris en version électronique, sont signalées comme telles. Ce travail n'a été soumis à aucun autre jury d'examen sous une forme identique ou similaire, que ce soit en France ou à l'étranger, à l'université ou dans une autre institution, par moi-même ou par autrui.

Le 15/07/2022

Signature :



Résumé

Ce mémoire a pour but d'étudier la variation des types de phonation entre individus, et la reconnaissance de celle-ci par divers prédicteurs tels que des auditeurs humains, un réseau de neurones et une mesure sur le signal acoustique. Cette variation est présente dans tous les aspects de la vie d'un locuteur, que les contraintes soient linguistiques, pathologiques ou encore culturelles.

Les résultats attendus étaient une variation perceptible des types de phonation selon les individus, ainsi qu'une meilleure performance du réseau de neurones par rapport aux deux autres prédicteurs.

Afin de les vérifier, nous avons cherché à voir comment chaque prédicteur classifiait les stimuli de notre corpus. Ceux-ci étaient issus du corpus PTSVOX ([Chanclu et al., 2020](#)), produits par vingt-quatre locuteurs (douze femmes et douze hommes) en parole lue et spontanée.

Pour les auditeurs humains, nous avons préparé un test de perception à l'aide de PsyToolKit ([Stoet \(2010\)](#), [Stoet \(2017\)](#)) : les auditeurs évaluaient le type de phonation de chaque stimulus, présenté dans un ordre aléatoire, et notaient leur réponse sur une échelle de Likert. Les auditeurs contactés pour l'étude étaient des spécialistes de la parole (chercheurs en phonétique, phonologie ou prosodie, orthophonistes). L'évaluation de la performance du réseau de neurones a été effectuée dans [Chanclu et al. \(2021\)](#). Quant aux mesures acoustiques, nous avons récolté les résultats de la mesure $h_1 - h_2$ opérée avec différents scripts Praat (le script de Styler *NasalityAutoMeasure* en modes automatique et manuel, et un script entièrement manuel que nous avons écrit).

D'après nos résultats, le réseau de neurones est bien celui ayant le mieux reconnu les types de phonation des locuteurs tels qu'ils étaient annotés dans le corpus. Il est suivi de près par les participants au test de perception. Cependant, les mesures acoustiques ont de mauvais résultats, en particulier pour la voix craquée. Cela est d'une part dû à des problèmes de détection de la f_0 — très observables pour les voix craquées de par leur irrégularité qui les caractérise — mais aussi car les scripts ne prennent pas en compte tout le stimulus comme le font les autres prédicteurs, mais seulement une courte portion du signal.

Les résultats montrent aussi que selon le locuteur, les auditeurs perçoivent un stimulus plus ou moins craqué ou soufflé. Cela montre que les types de phonation produits par les locuteurs possèdent une variabilité que des auditeurs peuvent percevoir.

Il est donc possible que certains individus possèdent une voix perceptiblement plus ou moins soufflée ou craquée, et une étude complémentaire pourrait s'intéresser à la reconnaissance et la mémorisation de ces locuteurs particuliers grâce à cet aspect de leur qualité de voix. Cela pourrait être utile dans le cas de la reconnaissance de locuteurs, notamment appliquée à des affaires criminelles.

Notre étude pose aussi la question de l'annotation du corpus, dont on voit en cherchant quelques stimuli dans les enregistrements, que les types de phonation produits changent dans le temps et sont parfois difficiles à identifier et à annoter correctement, surtout compte-tenu de la nature de la tâche.

Remerciements

Je tiens à remercier M.Cédric Gendrot pour avoir accepté de m'encadrer pour ce mémoire, ainsi que pour m'avoir aidée sur certains scripts et détails techniques de Praat.

Je remercie également M.Nicolas Audibert pour l'aide fournie dans l'analyse des résultats des participants au test de perception avec RStudio.

Table des matières

1	Introduction	1
2	État de l'art	2
2.1	Parole et langage	2
2.1.1	Appareil phonatoire	3
2.1.2	Qualité de voix	6
2.2	Types de phonation	7
2.2.1	Caractéristiques articulatoires	7
2.2.2	Caractéristiques acoustiques	8
2.2.3	Mesures acoustiques	9
2.2.4	Classification automatique	12
2.2.5	Paramètres paralinguistiques	14
2.2.6	Pathologies	15
2.2.7	Apport sociophonétique	16
2.2.8	Variations inter-locuteur	20
2.2.9	Reconnaissance du locuteur	21
3	Objectifs et hypothèses	25
4	Méthode	27
4.1	Extraction des stimuli	28
4.1.1	Résumé	31
4.2	Mise en place du test de perception	31
4.2.1	Résumé	34
4.3	Choix des mesures acoustiques	34
4.3.1	Résumé	37
5	Résultats	38
5.1	Performances des mesures	38
5.1.1	Mesures acoustiques	38
5.1.2	Résultats du test de perception	41

5.1.3	Prédictions du réseau de neurones	42
5.1.4	Comparaison des performances	44
5.1.5	Résumé	45
5.2	Analyse fine des stimuli	45
5.3	Évaluation des types de phonation des locuteurs	46
5.3.1	Résumé	48
6	Discussion	49
6.1	Critiques	51
7	Conclusion	53
8	Bibliographie	54
9	Annexe	59

1 Introduction

L'arrivée des nouvelles technologies a permis des qualités d'enregistrement bien supérieures, même pour des appareils tels que des téléphones portables. Pour la police scientifique, il s'agit d'autant de pièces à conviction potentielles pouvant servir dans le cadre d'enquêtes judiciaires. L'analyse de ces enregistrements requiert des compétences particulières en analyse du signal, et une formation en phonétique peut parfois être avantageuse afin de déterminer les caractéristiques acoustiques d'un individu.

Parmi celles-ci, nous nous intéresserons au type de phonation comme vecteur d'informations sur le locuteur, en essayant de déterminer si c'est une information importante et robuste sur celui-ci. La méthode pour évaluer le type de phonation d'un locuteur sera aussi discutée.

Nous aborderons d'abord la littérature présente sur le sujet de la parole et des types de phonation, puis énoncerons les hypothèses que nous testerons au cours de nos expériences et le protocole à suivre. Les résultats seront ensuite décrits et discutés, et nous conclurons sur la validation ou non des hypothèses formulées.

Tous les scripts cités, ainsi que les résultats obtenus et les stimuli utilisés, sont accessibles à l'adresse https://github.com/C-Millot/memoire_m1. Les liens exacts sont également fournis dans ce document.

2 État de l'art

Dans cette partie, nous allons définir les termes et le contexte importants pour comprendre les notions abordées durant notre étude. Nous évoquerons dans un premier temps les pré-requis phonétiques tels que le niveau glottal du conduit vocal et la qualité de voix, afin de définir ensuite les types de phonation sous divers angles de recherche, et ce qui peut provoquer de la variation entre les locuteurs les concernant.

2.1 Parole et langage

La parole est définie par le TLFi comme étant la « faculté d'exprimer et de communiquer la pensée au moyen du système des sons du langage articulé émis par les organes phonateurs. » (ATILF - CNRS & Université de Lorraine, sd). Théorisée dès Aristote au IV^{ème} siècle av. J.C., elle fait partie des formes de communication privilégiées dans les sociétés humaines du fait que la plupart des Hommes possèdent à la naissance les conditions nécessaires à son exaction — contrairement à l'écriture qui nécessite un support physique et une capacité d'abstraction supplémentaire.

D'après la définition du TLFi, nous pouvons remarquer certains axes sur lesquels s'articule cette citation.

La parole ne peut être dissociée de la notion de langage, en effet les sons produits pour parler sont régis par une langue : ils forment des signifiants qui servent à désigner les choses/concepts nécessaires à notre communication.

Dans certaines littératures, la parole est aussi présentée comme traduisant la pensée — une idée sans doute popularisée par Aristote, pour qui « *seul parmi les animaux l'homme a un langage. Certes la voix est le signe du dououreux et de l'agréable, aussi la rencontre-t-on chez les animaux [...] Mais le langage existe en vue de manifester l'avantageux et le nuisible, et par suite aussi le juste et l'injuste.* » (Aristote (v JC), Livre I, chapitre 2, p. 91 – 92). Cette relation entre parole et pensée est souvent débattue, par exemple par Macé au XVII^{ème} siècle qui pensait que la pensée était mue et conditionnée par la parole.

Enfin, le dernier aspect fondamental de la parole est la production de sons à l'aide de l'appareil phonatoire. Celui-ci est divisé en différentes régions : les parties supraglottique, glottique et sous-glottique.

Denes et al. (1993) résument ces aspects dans leur livre *The Speech Chain*, dans lequel ils

présentent le schéma de la communication, visible en Figure 1. Celui-ci montre que le locuteur conçoit d'abord son message dans son esprit à l'aide de son cerveau, puis le formule à l'aide de son conduit vocal. Le message traverse l'air sous forme d'ondes acoustiques pour parvenir aux oreilles de l'auditeur, puis au cerveau pour que le message puisse être déchiffré et compris.

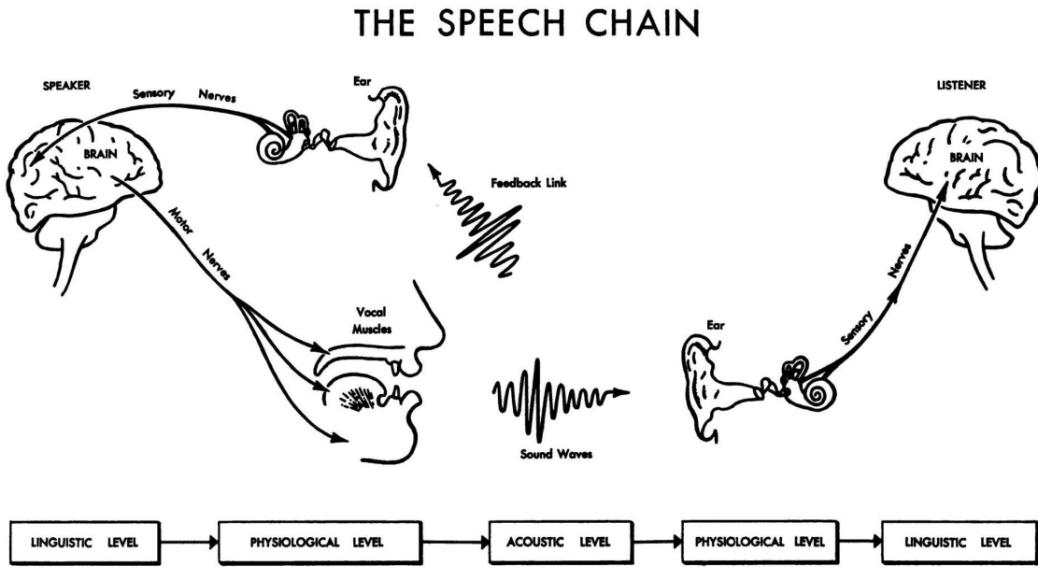


FIGURE 1 – Les différentes formes sous lesquelles un message produit peut exister, tandis qu'il progresse de l'esprit du locuteur à l'esprit de l'auditeur (d'après Denes et al. (1993), p. 17)

2.1.1 Appareil phonatoire

La région sous-glottique permet de créer un flux d'air, nécessaire à la production de certains sons appelés pulmoniques. En effet, l'air provient des poumons et est poussé dans le conduit vocal via l'action du diaphragme qui comprime ceux-ci. Il passe ensuite par la glotte, ou larynx. Celle-ci comporte des plis vocaux mus par les muscles laryngés tels que le muscle cricothyroïdien (voir Figure 2).

Ainsi, les plis vocaux peuvent se rapprocher l'un de l'autre ou s'éloigner, mais aussi s'allonger ou rétrécir, se raidir ou se relâcher... En se rapprochant assez l'un contre l'autre, ils peuvent fermer le conduit vocal et empêcher l'air de passer. Cependant, au fur et à mesure que la pression de l'air augmente sous la glotte tandis que le flux d'air est poussé par le diaphragme et le volume le contenant diminue, les plis vocaux ont de plus en plus de mal à rester accolés, et une fraction de l'air sous-glottique réussit à passer outre ceux-ci : c'est le principe de Bernoulli. La pression ayant diminué par cet écoulement, les plis se ré-accolent et le cycle continue. Chaque cycle accollement-

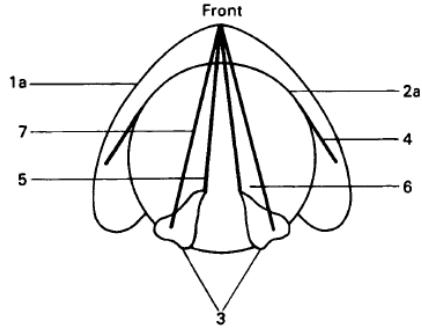


FIGURE 2 – Diagramme schématique de la position des muscles laryngés connectant le cartilage cricoïde au cartilage thyroïde, et les organes liés (d'après Laver (1980))

- | | |
|-----------------------------------|---------------------------------------|
| 1. Cartilage thyroïde | 4. Muscle cricothyroïdien |
| 1a. Bord extérieur de la thyroïde | 5. Frontière glottale du pli vocal |
| 2. Cartilage cricoïde | 6. Ventricule de Morgagni |
| 2a. Bord extérieur du cricoïde | 7. Frontière intérieure du ventricule |
| 3. Cartilages arytenoïdes | |

passage d'air est appelé un cycle glottique, et peut se produire en moyenne entre 100 et 250 fois par seconde — c'est la fréquence laryngée, aussi appelée fréquence fondamentale f_0 . L'écoulement régulier de l'air au travers des plis vocaux est à l'origine du voisement utilisé pour produire les voyelles et les consonnes voisées. Les consonnes sourdes sont produites avec les plis vocaux ouverts, ne perturbant pas l'écoulement de l'air.

Les cartilages arytenoïdes permettent aux plis vocaux de se rapprocher ou s'éloigner, tandis que les muscles cricothyroïdiens peuvent tendre longitudinalement les plis (Jiang et al., 2000) en faisant pivoter le cartilage thyroïde ; plus ceux-ci sont étirés, plus ils vibrent rapidement et plus la fréquence fondamentale est élevée.

Enfin, l'air rejoint la région supra-glottique du conduit vocal, composée du pharynx et des cavités nasale et buccale (langue, dents, palais dur et mou...). Le positionnement de ces articulateurs de la parole est déterminant dans la production d'un son de la parole, en effet ils ont tous la capacité de modifier le volume des cavités supra-laryngées, dans lesquelles l'air va résonner différemment. Le rapport de taille entre deux cavités engendre aussi une résonance supplémentaire, la résonance de Helmholtz. Les cibles de ces articulateurs supra-glottaux sont visibles sur la Figure 3, ainsi que la région glottale.

La théorie source-filtre permet d'expliquer les principes fondamentaux derrière l'utilisation du conduit vocal lors de la production de phonèmes (Fant, 1981) : on désigne par la source le larynx,

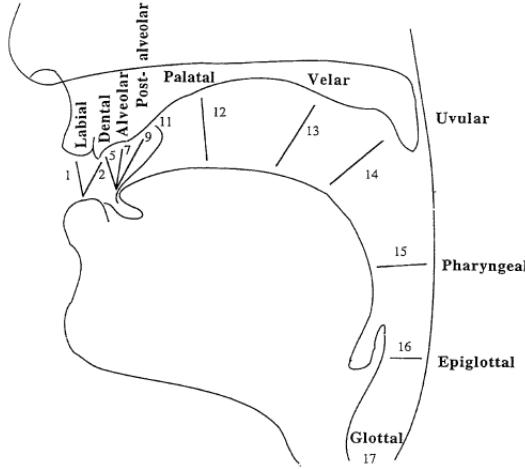


FIGURE 3 – Les neuf régions du conduit vocal pouvant être considérées comme des cibles pour les articulateurs mobiles. Les lignes numérotées montrent quelques-uns des dix-sept gestes articulatoires, incluant ceux dans la région glottale (d'après Ladefoged and Maddieson (1996), p. 13)

qui produit le bourdonnement brut du voisement.

Celui-ci est ensuite filtré par les articulateurs supra-laryngés, au fur et à mesure qu'il résonne dans les différentes cavités. Ces résonances seront à l'origine des formants, visibles sur les spectrogrammes de sons voisés. Il traduisent entre autres la taille relative des cavités (si la cavité antérieure — devant la langue — est plus grande, alors la fréquence du deuxième formant F2 sera plus élevée), comme on peut le voir sur la Figure 4.

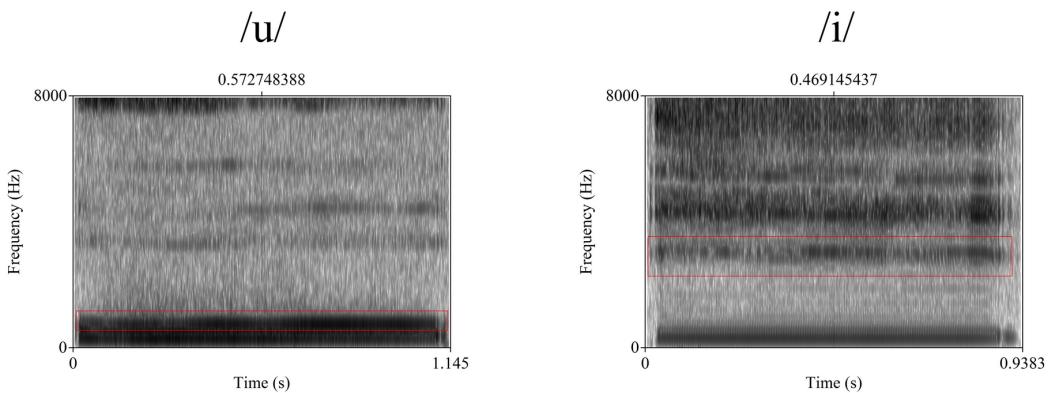


FIGURE 4 – Spectrogrammes montrant la hauteur de F2 pour des voyelles postérieure /u/ et antérieure /i/ extraits avec Praat (locuteur féminin)

Certaines consonnes comme les consonnes pulmoniques sourdes, ou les consonnes non pulmoniques comme les clicks, ne sont pas concernées par la source voisée du larynx. Cependant, le flux d'air est tout de même façonné par les articulateurs. Enfin, il existe des consonnes dont la source

n'est pas positionnée au niveau du larynx, mais au niveau supra-glottique : les fricatives ont une caractéristique bruitée importante de par la friction engendrée par le rapprochement important entre langue et palais.

2.1.2 Qualité de voix

La configuration des articulateurs et du larynx lors d'actes de parole peuvent déterminer la qualité de voix d'un locuteur.

Le concept de qualité de voix est plutôt large et permissif, ainsi la définition varie selon l'auteur. Par exemple, Barsties and De Bodt (2015) basent leur définition sur le caractère perçu de la qualité de voix, comme un phénomène perceptuel de la voix. Celui-ci est mal défini, mais pourrait être mesuré avec un test de perception, ou bien des mesures acoustiques sur des corrélats de qualités de voix.

Laver (1980), quant à lui, décrit les qualités de voix comme les « *caractéristiques auditives colorant la voix d'un individu* » (Laver (1980), p. 1). On pourrait dire que la qualité de voix d'un individu concerne le contenu non segmental de son discours, ainsi que les caractéristiques supra-segmentales propres à la langue parlée.

Enfin, Kreiman et al. (2003) définit les qualités de voix comme la façon « *dont les locuteurs projettent leur identité — leurs "caractéristiques physiques, psychologiques, and sociales" — au monde* » (Kreiman et al. (2003), p. 1), « qualité » étant synonyme de « timbre ». Selon eux, la difficulté à définir le terme de « qualité de voix » serait due au nombre important de fonctions servant la voix. Ainsi, des ajustements particuliers du conduit vocal supra-laryngé comme des lèvres plus arrondies pour toute production, ou bien une ouverture constante du velum laissant accès à la cavité nasale par le flux d'air (nasalité), sont des caractéristiques possibles de la qualité de voix d'un locuteur.

Cependant, au niveau laryngé, la façon dont les plis vocaux sont accolés peut aussi faire l'objet de variations. Nous avions vu que le rapprochement entre plis vocaux était à l'origine du voisement. Cet écoulement régulier du flux d'air est appelé voix modale. Cependant, certaines configurations où les plis sont plus ou moins accolés peuvent résulter en d'autres types de phonation.

Il est important de différencier le terme « type de phonation » du terme « qualité de voix », souvent utilisé par abus de langage pour décrire les mêmes phénomènes phonétiques. Ce serait car les types de phonation sont liés à la glotte, qui produit la voix, qu'ils sont souvent confondus avec le

concept-même de qualité de voix (Kreiman et al., 2003). Il est intéressant de noter que différentes caractéristiques de la qualité de voix (nasalité et type de phonation) ont les mêmes méthodes de mesure (pente spectrale), sur lesquelles nous reviendrons plus tard.

2.2 Types de phonation

Dans leur article *Phonation types : a cross-linguistic overview*, Gordon and Ladefoged (2001) présentent les types de phonation ainsi : « *Les Hommes peuvent contrôler leur glotte pour produire des sons de la parole avec, non seulement des vibrations voisées régulières sur une gamme de différentes hauteurs, mais aussi harsh [dures], soft [douces], craquées, soufflées et une variété d'autres types de phonation.* » (Gordon and Ladefoged (2001), p. 1, traduit de l'anglais).

Cet article a pour but de renseigner les paramètres, notamment acoustiques, des différents types de phonation. Comme de nombreux travaux décrivant les types de phonation, il a été en partie réalisé par Ladefoged, qui dès son PhD s'intéressait aux qualités des voyelles, puis a travaillé sur des mesures physiologiques de la respiration (Maddieson, 2007). Ce parcours l'a naturellement amené à évoquer les types de phonation.

Nous allons définir les types de phonation sous divers aspects — articulatoire, acoustique, mais aussi sociophonétique ou pathologique —, dans le but d'établir les facteurs de variation dans la qualité de voix d'un individu liée aux types de phonation (voix particulièrement rauque ou soufflée par exemple).

2.2.1 Caractéristiques articulatoires

Ainsi, il a présenté dans Ladefoged (1971) — un article présentant des concepts phonétiques indispensables à la bonne compréhension de la discipline —, une description physiologique des types de phonation en se basant sur l'ouverture relative entre les cartilages arytenoïdes (voir Figure 2), c'est-à-dire le degré de constriction des plis vocaux. La voix soufflée est celle pour laquelle les cartilages arytenoïdes sont les plus éloignés, et la voix craquée est celle pour laquelle ils sont au plus proche. La voix modale se situerait entre les deux. La Figure 5 illustre ce continuum.



FIGURE 5 – Continuum des types de phonation (Gordon and Ladefoged, 2001), d'après Ladefoged (1971)

Ladefoged and Maddieson (1996) ajoutent dans *The Sounds of the World's Language* les voix *slack* (détendue/relâchée) entre les voix soufflée et modale, et *harsh* (tendue, dure) entre les voix modale et craquée. De plus, peuvent être ajoutés aux extrêmes les productions sans voisement comme extension de la voix soufflée, et les occlusives glottales comme extension de la voix craquée. En effet, les plis vocaux ne sont pas du tout accolés pour les phonèmes sourds, et ils sont dans leur position la plus accolée pour une occlusive glottale.

Les voix *soft* et murmurée peuvent aussi s'inscrire dans le continuum de la voix soufflée. La voix *falsetto*, quant à elle, est caractérisée par un accollement fort des plis vocaux ainsi qu'un allongement des plis vocaux, donnant lieu à une voix plus aiguë (Esling, 1984).

Les corrélats articulatoires des types de phonation sont discutés dans ce livre : pour la voix modale, les cartilages arytenoïdes sont dans une position neutre — aucune force ne les écarte ni ne les rapproche. Les plis vocaux sont alors très proches, même s'ils ne sont pas forcément tout à fait accolés. Cette position neutre explique pourquoi ce type de phonation est particulièrement fréquent dans les langues du monde.

La voix craquée est caractérisée par un rapprochement important des cartilages arytenoïdes, ainsi qu'une tension plus grande dans les muscles du larynx. En particulier, si une partie des plis vocaux est trop proche des cartilages et est maintenue ainsi, elle ne vibrera pas. Ce type d'événement amène des vibrations irrégulières des plis vocaux car certaines parties ne vibreront pas en même temps. La voix tendue consiste en un rapprochement un peu moins important des plis vocaux. On ne peut caractériser ce rapprochement qu'en comparaison aux voix craquée et modale de la langue, car la voix tendue se situe dans un continuum entre les deux.

Les paramètres de la voix soufflée sont des cartilages arytenoïdes plus éloignés. Cette position entraîne un relâchement des plis vocaux, alors moins accolés donc plus ouverts. Ainsi, ils laissent passer davantage d'air. La voix relâchée est une version présente dans le continuum entre voix soufflée et voix modale, à l'image de la voix tendue.

2.2.2 Caractéristiques acoustiques

Concernant les caractéristiques acoustiques des qualités de voix, elles sont évoquées dans Gordon and Ladefoged (2001). Alors que la voix modale est caractérisée par un signal périodique, la voix craquée a des vibrations voisées irrégulières, généralement plus espacées dans le temps que pour la voix modale. La principale caractéristique de la voix soufflée est la présence importante de bruit

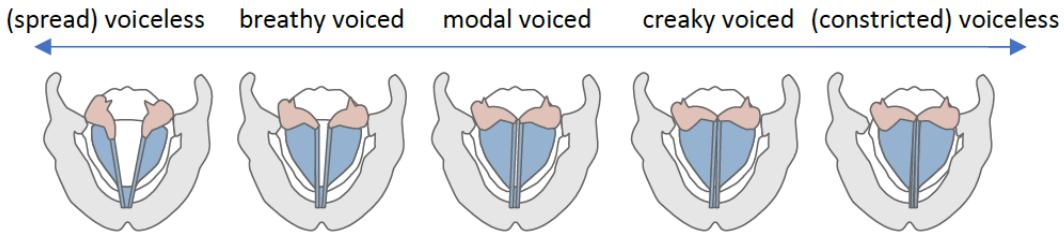


FIGURE 6 – Cinq types de phonation selon leur aperture glottale, de ouverte à fermée (d’après Wright et al. (2019))

dans le signal. Ainsi, les voix craquées et soufflées ont des composantes irrégulières, mais de nature différente.

Sur la Figure 7, on voit que le signal de la voix modale est régulier et périodique. Celui de la voix soufflée possède beaucoup de bruit, qui se voit également sur le spectrogramme — les harmoniques se distinguent moins. La voix craquée est très irrégulière : les impulsions sont d’ampleurs et de durées très différentes. Sur le spectrogramme, on distingue plus vivement les stries verticales, créées par les vibrations des plis vocaux. Elles sont espacées irrégulièrement car les vibrations sont elles-mêmes irrégulières.

On peut se poser la question de la comparaison entre les paramètres articulatoires et acoustiques des types de phonation : si les paramètres articulatoires font l’objet d’un continuum, peut-on en dire autant de leur perception acoustique ? Les études ne sont pas concluantes : « *Il est possible qu'il existe des états quantaux de la glotte comme suggérés par Stevens, mais ils ne sont pas faciles à déterminer en pratique.* » (Ladefoged and Maddieson (1996), p. 55, traduit de l’anglais).

2.2.3 Mesures acoustiques

Les mesures acoustiques importantes pour les types de phonation sont d’abord le *jitter* et le *shimmer*, deux mesures qui caractérisent des perturbations locales dans le spectre et peuvent donc être utiles pour étudier la voix craquée. En particulier, le *jitter* caractérise les perturbations locales (entre cycles laryngés consécutifs) de la f_0 , et le *shimmer* les perturbations locales entre cycles laryngés d’*amplitude* du signal. On trouve également la mesure de la pente spectrale.

Comme illustré sur la Figure 8, on voit en effet que la pente du spectre d’une voyelle craquée est montante sur ses deux premiers harmoniques, alors que celle d’une voyelle soufflée est descendante. Ainsi, si l’on soustrait à l’intensité du premier harmonique la deuxième et que l’on obtient la mesure

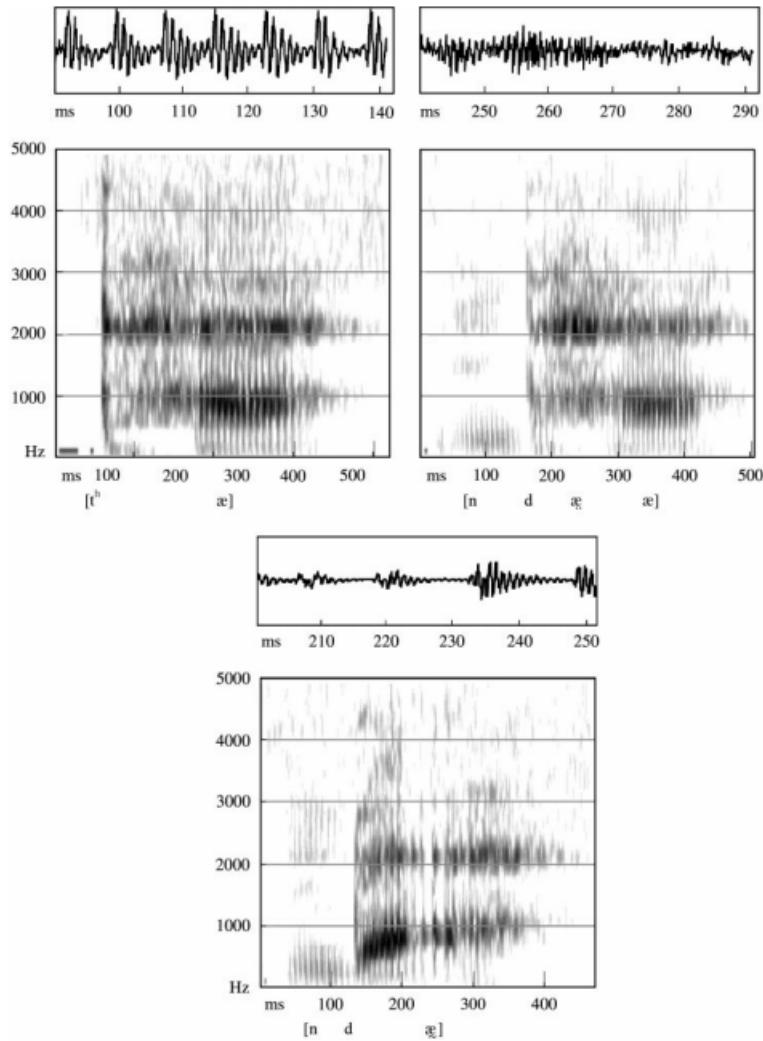


FIGURE 7 – Spectrogrammes et signaux de voyelles modale, soufflée et craquée des mots newar /nt^haé/ "graine", /ndaé/ "cheval", et /ndaé/ "fesses" (d'après Gordon and Ladefoged (2001), p. 390)

$h_1 - h_2$, le résultat sera plus grand pour une voix soufflée qu'une voix craquée, et sera modéré pour une voix modale.

Cette mesures pourraient être liées à des réalisation articulatoires — d'après Holmberg et al. (1995), la différence $h_1 - h_2$ est corrélée au pourcentage d'ouverture de la glotte durant un cycle glottal : plus la glotte est ouverte longtemps, moins l'amplitude de h_2 est grande comparée à celle de h_1 . Cela peut être appliqué à la voix soufflée, dont la production nécessite une ouverture plus importante des plis vocaux. À l'inverse, une voix craquée aura un quotient d'ouverture moins important, et donc un plus grand rapport entre h_1 et h_2 . Stevens (1977) observe que la brutalité de la fermeture glottale est corrélée avec une pente spectrale plus élevée (une différence plus grande entre

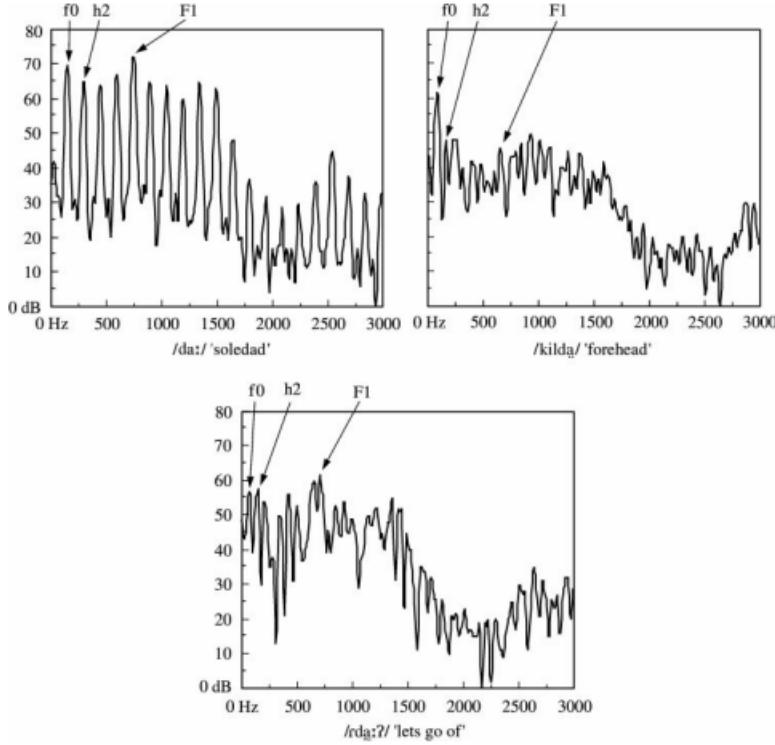


FIGURE 8 – Spectres en transformée de Fourier rapide d'un /a/ modal, soufflé et craqué dans les mots de zapotèque de San Lucas Quiavini /da:/ "Soledad", /kilda/ "front", et /rda:?:/ "lâche [quelque-chose]" (d'après Gordon and Ladefoged (2001), p. 398)

h_1 et h_2) : une fermeture glottale plus brusque serait alors liée à une voix craquée, alors qu'une fermeture plus douce serait liée à une voix soufflée.

L'article de Keating et al. (2010) cherche à pouvoir différencier les types de phonation intra ou inter-langues, et effectue à cet effet diverses mesures desquelles ils comparent l'efficacité : les mesures acoustiques $h_1 - h_2$; $h_2 - h_4$ (multiple de $h_1 - h_2$); $h_1 - A_1$; $h_1 - A_2$; $h_1 - A_3$; la prominence cepstrale; l'énergie; et une mesure articulatoire avec un électroglottographe, appareil permettant de mesurer le taux d'accolement des plis vocaux.

Les mesures acoustiques sont faites à l'aide de VoiceSauce (Shue et al., 2010) : c'est une implémentation à Matlab permettant de récupérer certaines mesures sur un enregistrement audio. Par exemple, il permet d'estimer la fréquence fondamentale avec différents programmes (STRAIGHT, the Snack Sound Toolkit et Praat. La possibilité de changer d'algorithme pour calculer la f_0 est intéressante car celle-ci peut parfois être difficile à détecter selon l'algorithme — des sauts d'octave peuvent avoir lieu). À partir de cette estimation, il peut ensuite calculer la position des harmoniques dans le spectre de l'extrait, puisqu'ils sont par définition des multiples de la f_0 . Les quatre premiers

formants, ainsi que les trois premiers anti-formants peuvent aussi être calculés.

Les résultats de l'article de Keating et al. (2010) montrent que seule la mesure $h_1 - h_2$ fonctionne pour différencier tous les types de phonation dans chaque langue, et entre chaque langue. Les mesures de type $h_1 - A_n$, les mesures d'énergie, et la prominence cepstrale, fonctionnent parfaitement mais seulement pour certaines langues. La validité ou non des mesures n'est pas influencée par le sexe ou la présence de tons dans la langue.

2.2.4 Classification automatique

Il existe également des techniques de classification automatique d'enregistrements audio selon leur type de phonation. L'avantage de mesures automatiques réside en une annotation moins sujette aux instincts des annotateurs humains qui ne sont pas forcément d'accord sur le seuil d'irrégularité à atteindre pour être craqué, et éventuellement plus rapide.

Chanclu et al. (2021) construisent par exemple un réseau de neurones dont la partie classificatrice distingue d'abord les extraits selon s'ils ont un type de phonation modal ou non, puis si le type de phonation non modal est craqué ou soufflé. Le système est testé sur un corpus français de voyelles prépausales produites en parole spontanée ou lue et annotées en type de phonation par un expert, et ses résultats sont comparés à ceux d'un système de référence utilisant trente MFCC — ce sont des coefficients représentant des caractéristiques acoustiques du signal permettant de rendre de la qualité de voix d'un individu, son timbre, dont son type de phonation.

Les résultats montrent que le réseau de neurones a un meilleur taux de classification correcte que le système de référence. De plus, les résultats sont meilleurs pour la classification entre les deux types de phonation non modaux, que ceux entre les types de phonation modal et non modal, comme une architecture en cascade — il est plus facile de classifier entre des catégories plus précises que globales.

Ce système est ingénieux car il permet de classifier les extraits craqués efficacement. En effet, la voix craquée, de par son irrégularité, propose un défi pour les systèmes de classification.

Vishnubhotla and Espy-Wilson (2006) proposent en solution une adaptation algorithmique du DéTECTeur APP (*Aperiodicity, Periodicity and pitch*). Celui-ci estime la proportion d'apériodicité de périodicité et de hauteur de la voix dans un signal de parole. Cependant, ce système seul ne peut différencier les phonèmes craqués, en effet les fricatives et les phonèmes soufflés ont aussi une composante apériodique. Pour filtrer les fricatives, on délaisse tout ce qui comporte de l'apériodicité

au-dessus de 3000Hz, en effet la voix craquée a peu d'intensité dans les hautes fréquences. Pour la voix soufflée, on filtre les signaux avec du voisement entre 0 et 1000Hz : les phonèmes craqués ont une phonation irrégulière à toutes les fréquences.

Les résultats sont bons, mais comportent une part de vrais négatifs dus aux occlusives de certains locuteurs.

Le système de Drugman et al. (2020) est plus simple et utilise des mesures acoustiques telles que $h_2 - h_1$ (qui donnera l'indice d'une pente spectrale inversée : une mesure positive indiquera la présence de voix craquée) ainsi que la prominence des pics spectraux — trouver les pics les plus hauts et calculer la différence d'amplitude entre chaque —, et des mesures présentées par Ishi et al. (2007). En mêlant ces mesures à un réseau neuronal, les résultats obtenus sont satisfaisants et font apparaître plusieurs types d'excitation spectrale due à la voix craquée.

Il existe certains algorithmes de détection de la voix craquée comme l'algorithme de Détection de Craquement, utilisé sous Matlab et qui permet de donner la probabilité de la présence de craquement selon un seuil, à partir d'un réseau de neurones évaluant certaines mesures acoustiques. Le paramétrage de ce seuil est important afin de garantir une bonne estimation des probabilités, comme le font White et al. (2021) afin de reconnaître la voix craquée de femmes australiennes ; ils montrent que le seuil optimal change selon la population étudiée, en effet le seuil optimal n'est pas le même pour des locutrices de l'anglais australiens et étasuniennes. Un autre exemple de réseau neuronal adapté à la reconnaissance de la voix craquée est DeepFry (Chernyak et al., 2022).

Kane and Gobl (2011) utilisent les ondelettes, une alternative aux sons purs comme description des sons de la parole, afin de différencier les voix craquée et soufflée. Les ondelettes ont l'avantage d'être robustes face aux enregistrements bruités. Après conversion du signal en ondelettes, on peut en tirer les maxima locaux à la manière des pics de Drugman et al. (2020) et effectuer une régression linéaire afin d'obtenir la pente du signal en un point donné. Cette pente est plus faible pour une voix soufflée, et plus forte pour une voix craquée.

Les résultats montrent que la pente permet de distinguer clairement chaque type de phonation, et permettent de mettre en valeur le concept d'ondelettes, encore peu usité en phonétique par rapport aux sons purs.

Il est intéressant de constater dans cette partie la diversité des mesures automatiques, allant d'algorithmes utilisant des mesures acoustiques $h_2 - h_1$ faisables à la main, à des réseaux de neurones incluant ou non ces dernières.

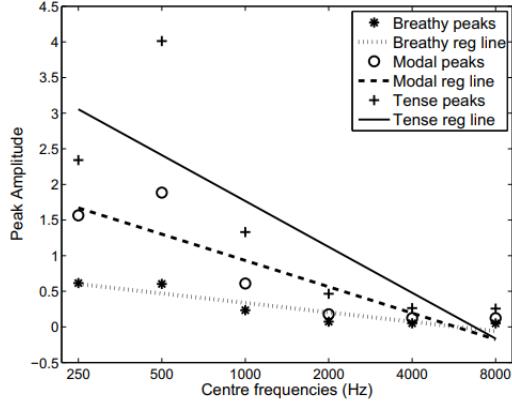


FIGURE 9 – Amplitude des pics d’ondelettes avec régression linéaire pour le milieu d’une voyelle /o/ produite par un locuteur dans des voix soufflée, modale et craquée (d’après Kane and Gobl (2011), p. 178)

2.2.5 Paramètres paralinguistiques

Les types de phonation ne sont pas des phénomènes linguistiques isolés phonétiquement, et peuvent apparaître en dehors de la volonté du locuteur — certains contextes phonétiques favorisent la production d’un type de phonation en particulier.

Par exemple, Esposito and Khan (2020) qui se proposent d’actualiser l’article *Phonation types : a cross-linguistic overview* de Gordon and Ladefoged (2001) rapportent que la coarticulation et certains processus phonologiques associés comme la coalescence peuvent provoquer des types de phonation. Notamment, le contexte /V?V/ donne par coalescence naissance à une voyelle craquée [V] en k’iche’ (Baird, 2011).

On trouve également en anglais des phénomènes de coalescence tels que la réalisation du contexte /hV/ en [V] (Ladefoged, 1983).

Enfin, la qualité de la voyelle peut aussi déterminer le type de phonation : en mong leng, la voyelle /a/ aurait une composante plus soufflée que /i/ ou /u/ (Andruski and Ratliff, 2000).

La structure prosodique de la langue peut aussi avoir une influence sur le type de phonation : Epstein (2002) relate les effets de la prominence d’un mot — en anglais — sur le type de phonation qui est produit, notamment pour les mots prominentes en début et en fin de phrase. Les résultats montrent que les mots prominentes ont un type de phonation plus tendu que les autres. Les syllabes en début et en fin de phrase sont souvent produites avec une voix craquée ou glottalisées. Enfin, une f_0 basse entraîne également souvent la production d’une voix craquée.

Les qualités de voix en général sont impliquées dans le découpage d'énoncés en groupes accentuels : Smith (2002) dit des énoncés déclaratifs et interrogatifs en français : « *Les questions ont un allongement final plus important, marquant la finalité en termes de durée. Les déclaratives, en revanche, ont une voyelle finale moins périodique et pour certains locuteurs, ont plus de chances d'être dévoisées, marquant la finalité en termes de qualité de voix.* » (Smith (2002), p. 166).

La conclusion de l'auteure est que l'apériodicité et le dévoisement seraient des marqueurs finaux car les voyelles non finales sont bien plus périodiques et bien mieux voisées.

2.2.6 Pathologies

Certains types de phonation peuvent être causés par des pathologies touchant la glotte.

Cela peut être le cas chez l'enfant suite à l'opération d'une sténose laryngotrachéale particulièrement grave, dont les symptômes sont un rétrécissement anormal important de la trachée jusqu'à la glotte, empêchant fortement la respiration.

L'opération consiste en une trachéotomie, c'est-à-dire une ouverture dans la gorge, afin de momentanément redonner au patient une source d'air. Celle-ci doit cependant être suivie de soins tels qu'une reconstruction laryngotrachéale (agrandir la trachée à l'aide de cartilages trouvés sur le patient) et une resection cricotrachéale (éliminer le morceau trop exigu de trachée et recoudre ensemble le larynx et la trachée), souvent accompagnés à terme par des changements dans la qualité de la voix.

Pullens et al. (2017) étudient ces changements via un questionnaire adressé aux parents des enfants opérés (l'Inventaire du Handicap de la Voix pédiatrique), ainsi que l'Indice de Sévérité de la Dysphonie, mesurant la fréquence fondamental la plus haute, la plus basse intensité, la durée de phonation maximale et le jitter : cet index peut ainsi estimer la qualité de la voix de l'individu, un résultat négatif indiquant une voix anormalement enrouée. Il est rapporté que, sur trente-huit patients âgés de moins de 18 ans, l'Indice de Sévérité de la Dysphonie médian était de -0,03, et pouvait aller de -6,57 à 5,78. Le score médian négatif montre bien que l'opération laisse des séquelles au larynx, et amène la production d'une voix tendue voire craquée.

Une autre cause pathologique de certains types de phonation sont les nodules sur les plis vocaux. Il s'agit de pseudo-tumeurs bénignes apparaissant sur la partie antérieure, au premier tiers, des plis vocaux (Lancer et al., 1988). Ils peuvent être bilatéraux — apparaître des deux côtés — symétrique

ou non, mais aussi unilatéraux. Une des conséquences de ces nodules est une voix plus soufflée, en effet les plis vocaux n'arrivent plus à s'accorder suffisamment pour maintenir une voix modale.

L'âge peut aussi jouer un rôle sur les types de phonation : Tarafder et al. (2012) rapporte une possibilité d'ossification et de calcification des cartilages laryngés, ainsi que une atrophie et une dégénérence des muscles associés au fur et à mesure du vieillissement. Cela entraîne des mouvements plus restreints des cartilages arytenoïdes, et donc des plis vocaux moins accolés durant la vibration. Ils peuvent aussi être moins souples de par une déshydratation, de ce fait la vibration peut être très irrégulière. Les auteurs recommandent ainsi une hydratation importante aux personnes âgées, ainsi qu'un entretien des plis vocaux en effectuant des vocalises, sans crier, et ne pas fumer.

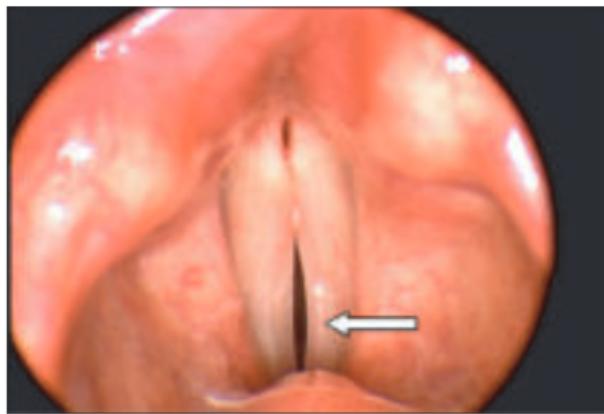


FIGURE 10 – Écart entre les plis vocaux dû à l'âge, les empêchant de se fermer totalement (d'après Tarafder et al. (2012), p. 2)

Enfin, l'âge mais aussi des comorbidités (hypertension, diabète...) augmentent les chances d'une paralysie des plis vocaux après une intubation trachéale. Cette paralysie se caractérise souvent par une dysphonie (Kikura et al., 2007).

2.2.7 Appart sociophonétique

Jusqu'ici, nous avons seulement vu comment les types de phonation peuvent être produits et leurs causes physiologiques, mais il peut être intéressant de voir comment ils sont perçus culturellement parlant.

Grâce au domaine de la sociophonétique, on peut décrire et étudier la signification sociale de productions sonores et leur variation, afin de « considérer la variation sociale et la production/perception de la parole ensemble » (Hay and Drager (2007), p. 90, traduit de l'anglais).

L'intérêt dans cette démarche est double : il s'agit de replacer la phonétique dans son contexte social afin de mieux comprendre les tenants sous-jacents des variations dans les productions orales, et comment les auditeurs perçoivent et normalisent les variations dans les discours ; mais aussi d'inclure les variations phonétiques comme vectrices de catégories sociales telles que l'âge, le genre, la classe sociale. Par exemple, la perception du /ɪ/ par des auditeurs néo-zélandais change selon que le contexte d'énonciation suggère l'Australie ou la Nouvelle-Zélande (locuteur supposément australien ou néo-zélandais, ou même une peluche de kangourou ou de kiwi dans la pièce) (Hay and MacLagan, 2006).

Nous pouvons déjà constater la présence des types de phonation dans de nombreux inventaires phonologiques, comme en !Xóõ (contraste présent pour les consonnes et les voyelles), en mazatèque (présent pour les voyelles) ou encore le newar (présent pour les consonnes uniquement) (Esposito and Khan, 2020), même sans aborder les caractéristiques sociophonétiques des types de phonation. Le mazatèque est d'ailleurs remarquable car il fait partie des rares langues à posséder un contraste de type de phonation à trois possibilités : voix modale, craquée et soufflée.

Il est toutefois important de préciser que les caractéristiques articulatoires et acoustiques des types de phonation peuvent différer selon la langue et le locuteur. Par exemple en mazatèque, pour la plupart des locuteurs le paramètre soufflé d'une voyelle se manifeste surtout pendant la première portion de celle-ci (Gordon and Ladefoged, 2001), mais en zapatèque c'est à la fin que la composante soufflée intervient (Esposito and Khan (2020), voir Figure 11). De plus, en hmong daw la composante soufflée d'une voyelle se trouve au début de celle-ci, alors que la composante craquée est au plus fort à sa fin (Esposito, 2005).

Benoist-Lucy and Pillot-Loiseau (2013) montrent dans leur article que la langue elle-même peut influer sur les types de phonation produits : dans leur expérience, elles étudient les types de phonation produits par des locuteurs américains en fonction qu'ils parlent français (en apprenants de L2) ou anglais. Les résultats montrent que lorsqu'ils parlent anglais, surtout en parole spontanée (par rapport à leurs productions en parole lue), les locuteurs ont une voix plus craquée que lorsqu'ils parlent français. Cette étude montre aussi que les apprenants d'une seconde langue n'occultent pas le choix du type de phonation et opèrent bien une réflexion à leur propos.

Nous pouvons ci-après en voir quelques exemples quant à l'influence de types de phonation dans la construction d'une image commune à une communauté.

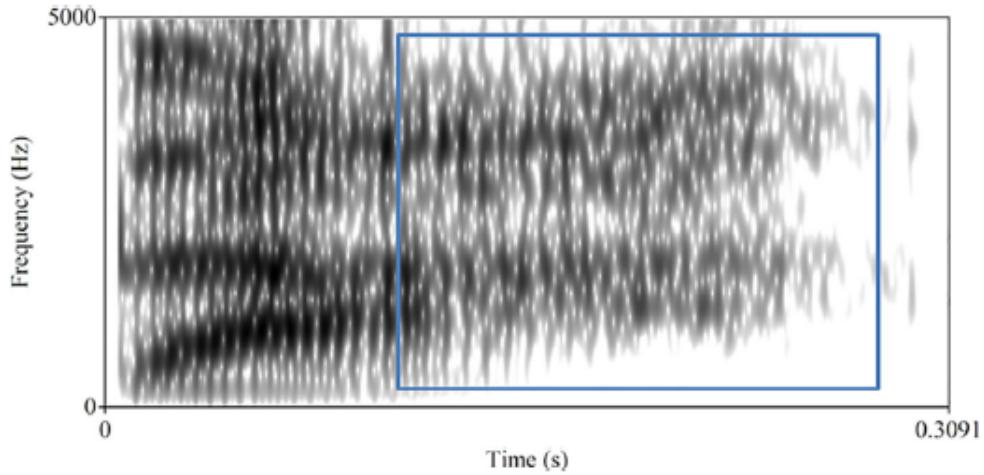


FIGURE 11 – Spectrogramme du mot zapotèque [dâ] « poudre » avec un rectangle illustrant la concentration de la voix soufflée à la fin de la voyelle (d'après Esposito and Khan (2020), p. 8)

Mendoza-Denton (2011) a publié une étude montrant l'importance de la voix craquée dans l'identité d'une gangster Chicano (mexicain habitant aux États-Unis, mais peut aussi inclure des personnes adoptant ce mode de vie) dur(e) à cuire (*hardcore*).

En effet, on remarque que des filles appartenant à ces gangs ont une utilisation prononcée de la voix craquée. Par exemple, « *you have to be down ~with yourself~ [...] you can't depend on ~people~* » (Mendoza-Denton (2011), p. 269, le texte entre tildes ~ est produit avec une voix craquée), ou encore une distinction entre « *guys* » pour signifier des personnes inconnues de la locutrice, à la différence de « *~dudes~* » qui désigne des hommes de son gang. Mendoza-Denton interprète la production de voix craquée comme intentionnelle et permettant de mettre une emphase sur des choses importantes aux yeux de la locutrice, comme animée par des sentiments de douleur, de tension. La voix craquée permettrait de faire transparaître ces émotions sans avoir l'air sentimentale, pour rester "dure à cuire". L'une des choses les plus importantes aux yeux des filles de gangs est effectivement de savoir se contrôler, d'avoir l'air *hardcore* ("dure de cœur").

Une autre utilisation de la voix craquée peut être vue chez les jeunes femmes américaines — le *vocal fry* : il s'agit pour ces locutrices d'utiliser la voix craquée de façon régulière, non pas dans le but de mettre une emphase sur un élément lexical (Gibson, 2017), mais plutôt, d'après Greer and Winters (2015), afin de renverser les attentes en utilisant un type de phonation davantage apprécié chez les hommes.

L'expérience de perception de Greer et Winters montre effectivement que chez un homme, la voix

craquée est plus appréciée et considérée comme « cool », alors que chez la femme elle renvoie à une personne peu fiable ; on lui préfère la voix soufflée. Les auteurs pensent que l'utilisation de ce type de phonation pourrait être une tentative pour les femmes de se démarquer (une voix craquée étant peu attendue d'une femme, et donc plus remarquable) et d'assimiler des caractéristiques traditionnellement masculines comme l'autoritarisme afin d'atteindre un statut social plus proche des hommes. Cela n'est cependant pas apprécié dans tous les milieux : alors que certaines études montrent que des femmes utilisant le *vocal fry* seraient moins susceptibles d'être embauchées (Anderson et al., 2014), le milieu universitaire estime davantage cette pratique (Yuasa, 2010).

Au Japon, une des performances de la féminité, notamment dans le doublage et le domaine professionnel de la voix, passe par ce que Starr and Greene (2006) appellent la *sweet voice*.

Celle-ci est caractérisée par une voix douce, presque chantante, produite avec une voix de tête et un peu de voix soufflée. Couplée avec un registre de langage propre au féminin dans la langue japonaise (utilisé seulement dans certaines contextes sociaux), elle produit l'effet de « *femmes plus âgées, dans des positions d'autorité féminine traditionnelle [...] bien élevées* » (Starr and Greene (2006), p. 17, traduit de l'anglais), et est donc utilisée pour doubler des personnages au rôle maternelle ou de grande sœur par exemple. Elle peut aussi être utilisée dans le cadre d'annonces publiques, comme dans une gare ou un métro. Il est intéressant de noter que cette voix n'est faisable que par des professionnels tels que des comédiens de doublage, de par la nécessité d'être en voix de tête ; nous avons donc affaire à la création d'un imaginaire collectif.

Dans Podesva (2006), Podesva étudie le lien entre l'utilisation de ce type de phonation chez les hommes et la performance de l'homosexualité. En étudiant les productions de 3 hommes homosexuels, l'auteur rapporte que le mode falsetto est en effet employé, bien plus que par des hommes hétérosexuels, et plus souvent dans la vie intime que professionnelle. L'intérêt d'utiliser ce type de phonation serait de marquer l'expressivité des locuteurs.

L'auteur note toutefois qu'il n'existe pas de manière de "parler gay", mais plutôt que le fait d'être expressif est interprété comme le signe de l'homosexualité du locuteur. Celui-ci serait conscient de cette interprétation, et souhaiterait être expressif car, dans l'imaginaire collectif occidental, cela est considéré comme sortant de la norme masculine et est davantage réservé aux rôles féminins.

2.2.8 Variations inter-locuteur

D'après ce que nous avons vu jusqu'ici, les types de phonation peuvent être produit pour nombre de raisons, que ce soit à cause de paramètres paralinguistiques, du contexte ou de la classe sociale, ou encore de pathologies. En plus de ces facteurs de variation, il existe des variations de type de phonation entre locuteurs, et même pour un locuteur donné.

Pour reprendre l'article Podesva (2006), l'utilisation du mode falsetto est associée à l'expressivité pour tous les locuteurs, néanmoins chaque locuteur l'emploie pour marquer un comportement en particulier : pour un locuteur il s'agit d'avoir l'air drôle et d'encourager à faire la fête, et pour un autre c'est une marque d'opposition (affective) à ce qu'un autre dit.

Hanson and Chuang (1999) s'intéresse aux variations, intra-locuteur et inter-locuteurs, dues au *glottal chink* : il s'agit d'une ouverture postérieure des plis vocaux qui a plusieurs répercussions acoustiques — la largeur des formants (surtout F1) est plus importante à cause de la perte d'énergie générée par l'ouverture, une pente spectrale plus intense et davantage de bruit d'aspiration dans le signal. On voit effectivement sur la Figure 12 que, pour le même texte prononcé, les locuteurs masculins ont chacun un degré de bruit d'aspiration très différent, correspondant peut-être à un *glottal chink* plus ou moins important. Il existe de plus des différences biologiques entre les productions féminines et masculines. On voit sur la Figure 13 que le spectre féminin a une pente spectrale plus importante et plus de bruit dans les hautes fréquences.

Cela pourrait donc indiquer des différences biologiques entre personnes de sexe féminin et masculin, se manifestant notamment dans la parole via les conséquences d'un *glottal chink* plus prononcé chez les femmes. Il existe donc des variations glottales entre individus dans leurs productions orales, ainsi que des différences marquées entre femmes et hommes. Peut-être que c'est cette ouverture, provoquant un léger souffle dans les productions féminines, qui a rendu populaire l'idée que la voix soufflée est plus plaisante chez une femme qu'une voix craquée que nous avions vue dans la partie 2.2.7.

Cette variation entre individus se retrouve aussi dans la prominence des mots en prosodie. Nous avions vu que les mots prominent ont un type de phonation plus tendu que les autres en anglais, que ce soit pour une phrase déclarative ou interrogative. Cependant, cette prominence dépend de l'individu : « *les locuteurs utilisent la qualité de voix pour distinguer entre des mots prominent*

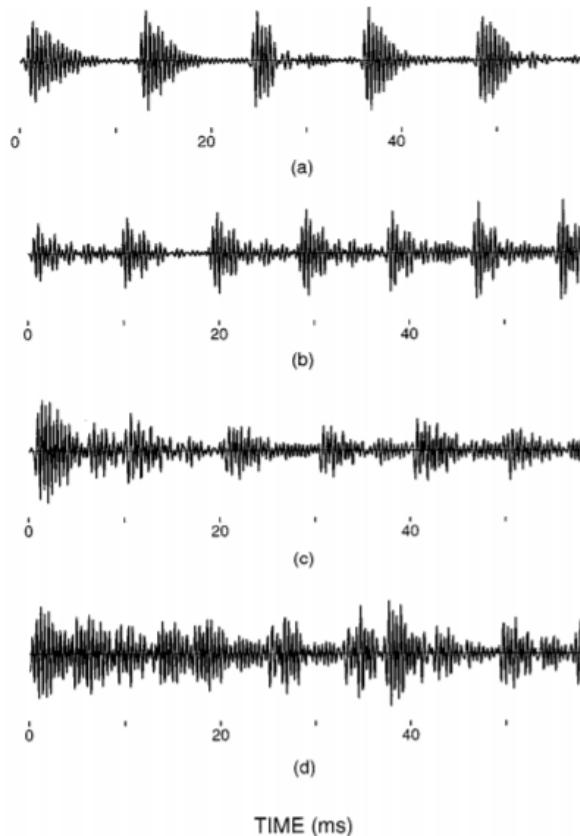


FIGURE 12 – Exemples de signaux sonores montrant des degrés de bruit d’aspiration différents selon le locuteur, masculin et différent à chaque fois (d’après Hanson and Chuang (1999), p. 1069)

ou non. Individuellement cependant, cela peut varier selon que l’effet soit plus fort pour les mots prominents dans des énoncés interrogatifs ou déclaratifs » (Epstein (2002), p. 94, traduit de l’anglais). On peut le voir sur la Figure 14 — les locuteurs B et S ont une voix plus tendue durant les mots prominents de leurs énoncés interrogatifs, tandis que pour le locuteur L cette prominence par le type de phonation est davantage utilisée pour les propositions déclaratives.

2.2.9 Reconnaissance du locuteur

Une des conséquences de la variation entre individus est l’hypothèse d’une voix propre à chaque locuteur, qui permettrait de l’identifier. C’est ce que font les humains lorsqu’ils entendent la voix d’une personne connue : ils essayent, avec plus ou moins de succès, de deviner de qui il s’agit sans voir la personne, et cela en faisant attention à leur qualité de voix particulière, dont notamment les types de phonation (Watt and Llamas, 2010). D’après Podesva and Callier (2015), les caractéristiques de la voix des personnes connues d’un individu sont encodées dans le cerveau sous forme de prototypes

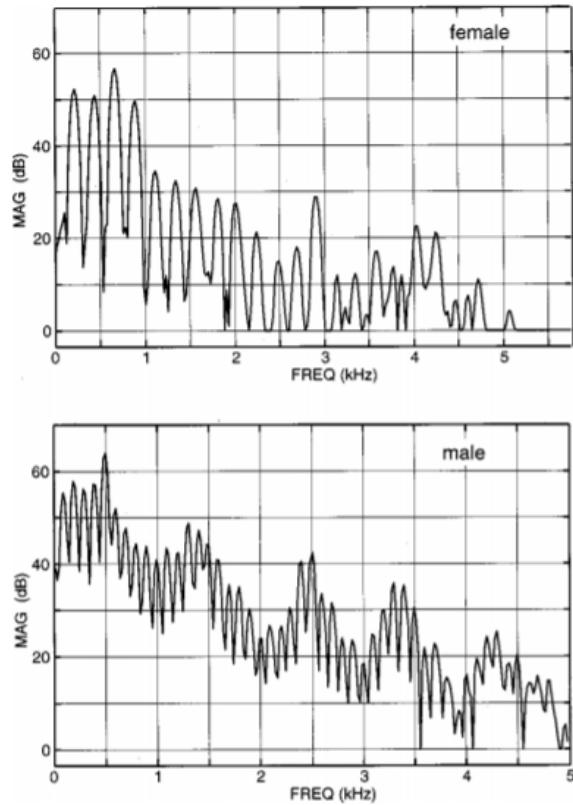


FIGURE 13 – Comparaison des spectres de la voyelle /ʌ(ʌ)/ pour des sujets femme et homme moyens (d'après [Hanson and Chuang \(1999\)](#), p. 1075)

vocaux à la manière dont on mémorise des visages, et non pas sous forme de traits acoustiques.

Cette supposée reconnaissance des locuteurs par leurs pairs pose alors la question de l'identification de locuteur d'après des enregistrements, en ajoutant à l'intuition des individus naïfs des mesures acoustiques précises. Cette identification serait particulièrement intéressante dans le cadre d'enquêtes policières, et de nombreuses recherches tentent d'améliorer la crédibilité des résultats de reconnaissance. [Jessen \(2008\)](#) décrit plusieurs types d'identification de locuteurs :

- on possède un enregistrement d'une personne inconnue dont on a des suspects (on peut donc comparer l'enregistrement à celui des suspects pour comparer) ;
- on a l'enregistrement d'un inconnu, mais aucun suspect. Il s'agira alors de faire du *profiling*, c'est-à-dire de récolter toutes les informations particulièrement remarquables de l'enregistrement afin de pouvoir reconnaître le suspect facilement ;
- il n'y a pas d'enregistrement, mais des témoins affirment reconnaître la voix d'un suspect.

L'utilisation d'enregistrements vocaux afin d'identifier des criminels commence à être mis à profit, même si ce type de preuve n'est presque jamais décisif pour l'arrestation d'un suspect — il

	Speaker B				Speaker L				Speaker S			
	EE	Lin	RK	OQ	EE	Lin	RK	OQ	EE	Lin	RK	OQ
All Sentence Types	X	X		X					X		X	X
Declaratives					X	X	X	X				
Interrogatives	X	X	X	X					X	X	X	X

FIGURE 14 – Paramètres utilisés pour indiquer une plus grande tension des plis vocaux pour des mots prominent, séparés par type de proposition et par locuteur. Les effets de tension considérables sont représentés par une croix dans la case, et des effets moins importants ont une croix en pointillés. Les paramètres sont : EE (intensité spectrale), RK (symétrie glottale), OQ (quotient ouvert) et Lin (linéarité spectrale)

ne s'agit souvent que de preuves concordantes qui font pencher la balance en faveur d'un suspect. Morrison (2009) précise que le modèle actuel de la *forensic phonetics* est celui d'une méthode quantitative donnant les probabilités que se trouvent des similitudes entre deux enregistrement. Les comparaisons sont maintenant souvent faites entièrement via des logiciels informatiques, sans intervention humaine. Néanmoins, des experts en phonétique peuvent toujours être appelés afin de cibler l'analyse sur un aspect phonétique ; par exemple, la diphongaison des voyelles.

L'un des exemples les plus connus est le *profiling* du faux Éventreur du Yorkshire ; à la suite d'une série de meurtres en Angleterre perpétrée par l'Éventreur du Yorkshire, la police britannique avait reçu une présumée confession audio du tueur avec des lettres. Contacté par la police, le phonéticien Ellis a réussi à établir le profile du tueur (déduire à un *mile* près son lieu de vie) d'après son accent, constitué de diphongaisons du /u:/ ou encore de la déletion du /h/ dans < having >. Ils ne purent néanmoins pas identifier de suspect habitant la zone décrite par Ellis et sans alibi pendant les meurtres, en effet ils apprirent après avoir trouvé le coupable que l'auteur des confessions n'était qu'un imposteur, grâce à des tests ADN. Ce fut onze ans après les fausses confessions que l'imposteur fut arrêté.

En plus d'être un exemple de *profiling* réussi, cette affaire montre que les enregistrements de suspects, même âgées dans le temps, peuvent être utiles pour identifier des locuteurs : Künzel (2007) a étudié l'impact possible de l'âge d'un enregistrement dans la perception/reconnaissance d'un locuteur par rapport à sa voix actuelle. Les résultats montrent que l'écart de temps entre les deux productions (âgée et récente) ne gêne pas la reconnaissance et que le locuteur reste identifiable : sa voix ne change pas suffisamment pour devenir une contrainte.

Il ne faut cependant pas oublier, comme le rappellent Watt and Llamas (2010), que l'empreinte

vocale — en référence à l’empreinte digitale — ne peut exister de par la nature de la parole : celle-ci n’est pas tangible, et seules les ondes et résonances résultant du conduit vocal nous parviennent, jamais le conduit lui-même.

Nous avions évoqué que les humains n’étaient pas toujours des sources fiables en matière de reconnaissance de voix. L’approche de Gerlach et al. (2020) pourrait alors se révéler intéressante dans le cadre judiciaire, afin d’avoir une méthode objective de comparaison de voix. Ceux-ci proposent de comparer les estimations de similarité entre voix par des auditeurs humains, et par un système de reconnaissance du locuteur (VOCALISE) utilisant des caractéristiques acoustiques telles que les formants ou la hauteur de la voix. Les estimations s’avèrent être très similaires, et les auteurs envisagent son utilisation lors de la sélection de plusieurs voix similaires à présenter à des victimes afin d’identifier un suspect. Toutefois, les estimations ne sont pas tout à fait identiques car des informations sociophonétiques et linguistiques peuvent aider les juges humains à faire mieux que le système de reconnaissance.

3 Objectifs et hypothèses

D'après ce que nous avons pu voir, la qualité de voix d'un locuteur peut-être en partie déterminée par un type de phonation prédominant chez celui-ci. Cela peut être dû à de nombreuses raisons, physiques comme culturelles : le vieillissement ou le sexe du locuteur, mais aussi la langue parlée ou l'appartenance à un certain groupe social.

Ainsi, le type de phonation peut être un indice dans la détermination des caractéristiques d'un locuteur, et il est possible que certains locuteurs aient un type de phonation plus prononcé qu'un autre, permettant son identification.

Notre premier objectif sera d'étudier si certains locuteurs possèdent un type de phonation particulier qui teinterait leurs productions vocales — si l'un de leurs types de phonation est plus proéminent et détectable par rapport à d'autres locuteurs. C'est là notre première hypothèse : certains locuteurs ont un type de phonation plus saillant et reconnaissable.

De plus, nous l'avons vu, les méthodes de classification des types de phonation sont diverses, et l'on utilise à présent autant des experts humains qualifiés, que de mesures sur le signal acoustique ou de classification automatique (par réseau de neurones) afin de distinguer les types de phonation présents dans un énoncé.

Notre autre objectif sera de comparer l'efficacité de prédiction de ces différents classificateurs, en observant lesquels permettent de prédire avec plus de précision le type de phonation. Il sera intéressant de voir à quel point les prédictions des classificateurs sont similaires, et si des groupes de prédictions similaires se forment entre certains. La deuxième hypothèse est donc que tous les prédicteurs n'auront pas les mêmes taux de réussite. D'après la littérature, les réseaux de neurones sont très compétents, ainsi que la mesure $h_1 - h_2$ qui est très populaire afin de calculer la pente spectrale. La qualité de l'évaluation par des auditeurs est quant à elle parfois critiquée, donc les résultats pourraient être moins bons. Il reste à voir si le réseau de neurones sera plus proche des auditeurs, ou des mesures acoustiques.

- Si notre hypothèse sur les types de phonation des locuteurs est avérée exacte, nous nous attendons à voir des taux de reconnaissance différents des types de phonation selon les locuteurs : certains pourraient avoir une voix soufflée particulièrement mieux reconnue que d'autres, par exemple.
- Si tous les prédicteurs n'ont pas les mêmes taux de réussite, cela se verra dans les résultats

— le plus probable serait qu'un réseau de neurones ait de très bons résultats, tandis que les auditeurs humains seraient un peu moins bons. Au sujet des mesures acoustiques, leur robustesse sera mise à l'épreuve sur des extraits de parole continue, donc il est difficile de prévoir leur réussite.

4 Méthode

Le corpus utilisé pour cette étude est le corpus PTSVOX (Chanclu et al., 2020), dont les types de phonation selon la voyelle et le mot sont annotés à l'aide de TextGrids sur Praat. Notre recherche s'intéressant à la comparaison entre perception d'auditeurs et classification d'un réseau de neurones sur les mêmes extraits, le corpus utilisé doit être le même que celui de Chanclu et al. (2021).

Le corpus PTSVOX (Chanclu et al., 2020) est composé d'enregistrements de 369 locuteurs et locutrices d'écoles de police, dont vingt-quatre (douze hommes et douze femmes) enregistrés plusieurs fois sur une période de plusieurs mois. Les âges des participants, langue maternelle et état de santé (cigarette, opérations dans la zone ORL...) sont renseignés. Les enregistrements, qui comptabilisent en tout plus de quatre-vingt heures, sont faits avec un micro ou un téléphone portable. En effet, la récolte des données s'articule autour de la question de la variabilité dans divers enregistrements d'un même locuteur, afin d'améliorer la reconnaissance et la comparaison de voix dans le cadre judiciaire. En constituant un corpus avec un grand nombre de locuteurs, de méthodes d'enregistrement et de types de parole (parole lue, et surtout parole spontanée), les auteurs espèrent obtenir des données adaptées à la comparaison de voix car les autres études s'en servent davantage dans un cadre commercial. Plus particulièrement, ce corpus a été établi afin d'être utilisé par un réseau neuronal (Chanclu et al., 2021).

Les enregistrements sont ensuite transcrits et les enregistrements micro sont annotés selon leur type de phonation. Seuls quatre types sont retenus : voix craquée, modale, soufflée et aspirée.

Ce sont ces enregistrements que nous utilisons pour constituer nos données. En effet, l'article précise que la qualité des enregistrements téléphone était plus mauvaise que celle des microphones et a posé problème à des systèmes de classification. De plus, nous concentrons notre collecte sur les données des vingt-quatre locuteurs récurrents dans le corpus, afin de pouvoir mieux étudier la variabilité intra-locuteur en plus de celle inter-locuteurs.

En tout, il y a 9786 voyelles annotées. Le Tableau 1 montre le nombre de voyelles par locuteur.

Comme pour nos résultats attendus, notre question de recherche est adressée sous deux angles : l'un orienté vers les mesures acoustiques sur les échantillons sonores prélevés dans le corpus, et l'autre vers la perception et l'identification des types de phonation de ces échantillons, par des participants. L'analyse acoustique nous permet aussi d'identifier les éventuels facteurs pertinents dans la perception des extraits sonores.

Locuteur	Nombre de voyelles produites
Locuteur 1	427
Locuteur 2	452
Locuteur 3	416
Locuteur 4	411
Locuteur 5	434
Locuteur 6	163
Locuteur 7	328
Locuteur 8	477
Locuteur 9	415
Locuteur 10	387
Locuteur 11	467
Locuteur 12	379
Locuteur 13	497
Locuteur 14	313
Locuteur 15	537
Locuteur 16	462
Locuteur 17	355
Locuteur 18	345
Locuteur 19	412
Locuteur 20	359
Locuteur 21	416
Locuteur 22	480
Locuteur 23	391
Locuteur 24	445
Total	9678

TABLE 1 – Nombre de voyelles produites par chaque locuteur dans le corpus PTSVOX

Le premier aspect de notre recherche a été de déterminer ces échantillons. De par notre intérêt pour la variation des types de phonation selon le locuteur, il était important de garder des extraits de tous les locuteurs, sans avoir un nombre de stimuli trop important pour autant.

4.1 Extraction des stimuli

Les stimuli recherchés étaient des voyelles positionnées avant une pause dans l'énoncé, dans le but d'avoir une longueur suffisamment importante pour percevoir correctement le type de phonation. En effet, le but n'était pas de piéger les participants en leur donnant des extraits difficiles. L'autre avantage d'une voyelle devant une pause est qu'elle est moins influencée par son contexte suivant.

Nous avons décidé de prendre les occurrences de deux voyelles différentes afin de ne pas créer de biais comme un type de phonation intrinsèque à une voyelle. Au delà de ce nombre nous aurions eu

trop de stimuli. La peur d'un nombre de stimuli trop important nous a également dissuadé d'ajouter une répétition pour vérifier si les réponses des participants étaient cohérentes.

Afin de déterminer les voyelles à utiliser, nous avons recherché celles comportant toutes les qualités de voix dans un même contexte, pour que la variation de celui-ci n'impacte pas la production par le locuteur. Cela se fit en deux étapes.

D'abord, nous avons créé le script Praat (https://github.com/C-Millot/memoire_m1/blob/main/tableau_occurrences_voyelles.praat) prenant tous les fichiers .wav et .TextGrid d'un dossier, repère toutes les voyelles de chaque fichier (selon la tier qu'on lui a indiquée) et enregistre dans un fichier le type de phonation qui lui est associé (on prend celui indiqué au début de la voyelle et aussi à la fin, car il y a parfois un changement de type de phonation durant une même voyelle : le type de phonation peut ne pas être stable), le phonème qui précède la voyelle (son contexte), la durée de la voyelle et l'indice temporel elle commence.

En ouvrant le fichier obtenu dans un tableur, nous avons étudié les voyelles ayant toutes les qualités de voix dans un certain contexte. Les contextes les plus présents étaient /t/ pour les voyelles /a/, /e/, /ə/, /i/, /u/, /ɔ/, et /s/ pour /e/ et /ə/.

Nous avons ensuite extrait ces voyelles pour identifier celles à utiliser dans nos analyses. Pour cela, nous avons utilisé un autre script Praat (https://github.com/C-Millot/memoire_m1/blob/main/extract_vowels.praat) dans lequel on précise la voyelle, le type de phonation et le contexte recherchés et on extrait les voyelles correspondantes, ainsi qu'un deuxième script qui extrait les mots dans lesquels sont ces voyelles (https://github.com/C-Millot/memoire_m1/blob/main/extract_words.praat), ce qui nous permettait une écoute plus globale des futurs stimuli. En étudiant les extractions, nous avons constaté que tous les locuteurs n'étaient pas représentés équitablement dans les fichiers. Notre but étant de privilégier la reconnaissance des types de phonation selon le locuteur, il était donc important qu'ils soient tous représentés.

Ainsi, discriminer le contexte des voyelles n'était pas possible sans sacrifier la représentation des locuteurs, et nous avons dû extraire les stimuli sans distinction de contexte.

Afin de continuer la sélection de nos stimuli, nous avons dû décider des voyelles à extraire, en effet nous recherchions deux voyelles distinctes pour avoir de la variation dans nos stimuli. Les voyelles les plus fréquentes dans nos fichiers étant /ə/ et /e/, nous les avons choisies. Il nous fallu également décider de la durée de nos stimuli ; le corpus comporte en effet des extraits allant

de seulement quelques dizaines de millisecondes, à plus de 400 millisecondes. D'après l'écoute des différents fichiers, nous avons jugé que, pour des extraits plus courts que 150 millisecondes, il était difficile de pouvoir correctement les identifier. Nous avons donc établi une plage de durée entre 150 et 200 millisecondes afin d'avoir des extraits identifiables, mais pas trop faciles non plus.

En extrayant cette fois des fichiers sonores selon les voyelles /ə/ et /e/, une durée entre 150 et 200 millisecondes et le type de phonation (https://github.com/C-Millot/memoire_m1/blob/main/extract_stimuli.praat), nous avons obtenu 24 fichiers (pour les vingt-quatre locuteurs) * 3 types de phonation * voyelle (/ə/ ou /e/ sans distinction), soit soixante-douze fichiers.

Si l'on essaie de faire des fichiers séparés selon la voyelle (trier entre /ə/ et /e/), on voit qu'il y a bien plus de résultats pour /ə/ que pour /e/, surtout pour les voix craquée et soufflée qui sont moins présentes chez /e/ particulièrement à partir du dix-neuvième locuteur.

Nous avons pu vérifier cela en regardant la nomenclature de nos fichiers, triés par locuteur/-voyelle/type de phonation : nous avons ainsi pu nous rendre rapidement compte de l'existence d'un fichier pour une certaine combinaison ou non. Le poids des fichiers pouvait aussi nous informer sur le nombre d'extraits pour chaque combinaison.

L'étape suivante fut de sélectionner un ou deux stimuli par locuteur et par qualité de voix. Pour effectuer ce dernier tri, nous nous sommes basés sur les résultats (https://github.com/C-Millot/memoire_m1/blob/main/CNN_craque.xlsx, https://github.com/C-Millot/memoire_m1/blob/main/CNN_modal.xlsx, https://github.com/C-Millot/memoire_m1/blob/main/CNN_souffle.xlsx) de Chanclu et al. (2021), en essayant d'avoir quelques occurrences où le réseau avait mal prédit le type de phonation ; cela ne fut pas toujours facile, car le réseau de neurones devine souvent le type de phonation avec une probabilité acceptable (supérieure à 0,80). La stabilité de la qualité de voix (annotée dans la TextGrid respective de chaque fichier son du corpus PTSVOX) joua aussi un rôle dans la sélection.

Au final, après recherche, nous n'avons pas pu récolter toutes les qualités de voix pour chaque locuteur, car nous avons fait primer la stabilité du type de phonation.

Nous avons ainsi extrait soixante-douze stimuli, ce qui nous parut être un nombre satisfaisant. Nous avons créé un tableau récapitulatif pour nos stimuli, renseignant le nom de chaque fichier correspondant à un stimulus différent, le fichier son du corpus contenant l'extrait sonore, le locuteur, l'indice temporel du point de départ de l'extrait dans ce fichier son, la durée du stimulus, la voyelle

produite, le type de phonation, et comment celui-ci est reconnu par le réseau de neurones. Cela nous a aidé à avoir des informations sur les stimuli afin de pouvoir expliquer nos résultats acoustiques et perceptifs (https://github.com/C-Millot/memoire_m1/blob/main/tableau_stimuli.xlsx), et nous y avons ajouté les taux de reconnaissance par participants après l'expérience. L'ensemble des stimuli est accessible ici : https://github.com/C-Millot/memoire_m1/tree/main/Stimuli.

4.1.1 Résumé

La récolte des stimuli s'est articulée autour de contraintes :

- un nombre de stimuli adapté à un test de perception ;
- avoir tous les locuteurs représentés ;
- deux voyelles ;
- une durée entre 150ms et 200ms ;
- trouver des stimuli bien ou mal reconnus par le réseau de neurones.

4.2 Mise en place du test de perception

Afin de tester la perception des types de phonation selon les locuteurs, nous avons utilisé la plateforme en ligne PsyToolKit ([Stoet \(2010\)](#), [Stoet \(2017\)](#)) afin d'y construire un test de perception.

PsyToolKit, créé par Stoet, est un outil en ligne permettant la mise en place de tests de perception et de questionnaires grâce à un langage de script développé par l'auteur. Destiné dans un premier temps à des études dans le domaine de la psychologie, ses usages se sont avérés adaptés à la recherche en phonétique et la plateforme est maintenant régulièrement utilisée à cet effet, sa gratuité étant un avantage considérable.

La plateforme permet de créer des questionnaires préalables au test afin de connaître certaines caractéristiques des locuteurs (âge, sexe...). Les possibilités de création du test sont nombreuses : des stimuli sonores, vidéo, textuels... Peuvent être joués, et les mesures vont de l'exactitude de la réponse au temps de réaction entre le stimuli et la réponse du participant. Les résultats sont ensuite mis à disposition en téléchargement sur la plateforme. Nous avons envoyé ce test à des personnes travaillant dans le milieu de la phonétique/orthophonie, à même de connaître les différents types de phonation et leurs corrélats acoustiques et spectrographiques.

Le but du test était d'étudier comment sont reconnus les différents types de phonation selon les locuteurs, et si cela correspond ou non à la reconnaissance du réseau de neurones. Un pré-test afin d'évaluer l'expérience (si elle est trop longue, mal expliquée...) fut envoyé à des étudiants en

sciences du langage, d'un niveau inférieur à un master.

Les stimuli utilisés étaient ceux extraits durant la phase de récolte des stimuli, auxquels nous avons ajouté six stimuli d'entraînement (la voyelle /a/, deux par type de phonation) afin que les participants soient à l'aise avec l'expérience.

Le test se déroulait ainsi : une présentation globale des enjeux du test et de l'auteur du mémoire, puis une partie Questionnaire (https://github.com/C-Millot/memoire_m1/blob/main/psytoolkit_survey.txt) où sont inscrites des informations sur le test — le passer dans le calme avec casque et souris/pad —, et les consignes suivantes :

« Vous allez entendre des voyelles de type /ə/ ou /e/ : votre but sera de choisir parmi trois types de phonation en indiquant votre degré de certitude. Vous pourrez répéter le stimulus autant de fois que nécessaire.

À tout moment il vous est possible de faire une pause, en effet vous avez un quart d'heure pour répondre à chaque essai, qui se fait généralement en moins de 15 secondes. ».

Venait ensuite la partie Expérience (https://github.com/C-Millot/memoire_m1/blob/main/psytoolkit_experiment.txt) : les participants cliquaient sur un carré rouge pour commencer le test de perception. Un spectrogramme du stimulus était affiché en haut ; en bas se trouvaient les boutons "Voix craquée", "Voix modale", "Voix soufflée" pour répondre et un bouton pour rejouer le son (voir Figure 15). Ensuite, une échelle de Likert apparaissait afin que les participants notent à quel point ils étaient sûrs de leur réponse entre 1 (peu sûr) et 5 (très sûr) (voir Figure 16).

Pendant la phase d'entraînement, la réponse à chaque stimulus était affichée sur l'écran après vote en indiquant si elle était bonne, et la réponse à l'échelle de Likert était aussi montrée. Ainsi, les participants recevaient un *feedback* et pouvaient mieux comprendre la passation du test.

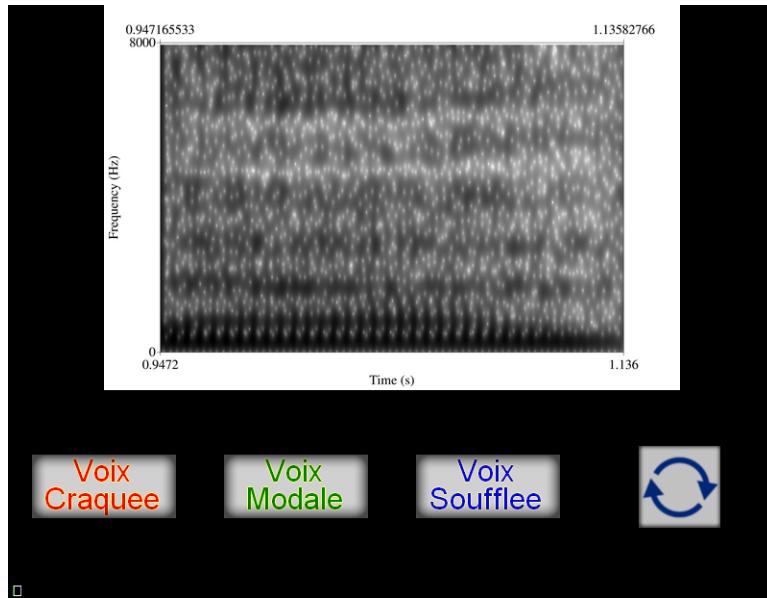


FIGURE 15 – Exemple de tâche d’identification de qualité de voix dans notre test de perception sur PsyToolKit

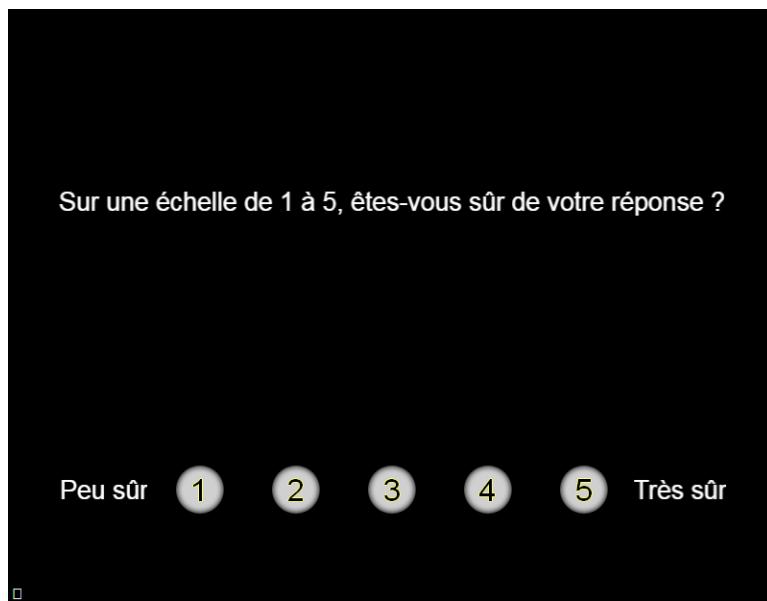


FIGURE 16 – Exemple d’échelle de Likert dans notre test de perception sur PsyToolKit

Le test a été partagé à une liste de personnes travaillant dans le domaine de la phonétique ou de l’orthophonie, francophones. Suite à la diffusion par l’une d’entre elles du test à des étudiants, nous avons ajouté au questionnaire les deux questions suivantes pour pouvoir filtrer les réponses d’éventuels étudiants trop peu qualifiés.

- Quel est votre degré d'expertise en phonétique/prosodie, ou orthophonie ? (de 1 = Aucune connaissance dans ce domaine à 7 = Pratique professionnelle)
- Comment avez-vous eu accès à ce test ?
 - J'ai reçu un lien directement de Carole Millot ;
 - Un de mes professeurs me l'a partagé ;
 - Un collègue me l'a partagé.

Cet ajout a nécessité quelques ajustements lors de l'évaluation des résultats.

Au final, nous avons recueilli les réponses de dix-neuf participants.

4.2.1 Résumé

Nous avons conçu le test de perception sur PsyToolKit ainsi :

- une tâche d'identification de la qualité de voix, avec spectrogramme du stimuli et entraînement avec *feedback* ;
- l'évaluation par le participant de sa réponse, sur une échelle de Likert ;
- 72 stimuli ;
- les participants (dix-neuf au total) connaissaient bien les types de phonation et leurs corrélats acoustiques.

4.3 Choix des mesures acoustiques

Nous avons effectué trois mesures acoustiques différentes sur nos données.

D'abord, des analyses automatisées à l'aide du script Praat *NasalityAutomeasure* de Styler (accessible ici : https://github.com/stylerw/styler_praat_scripts). Ce script permet d'obtenir de nombreuses mesures vocaliques liées à la nasalité, telles que la fréquence de formants, l'amplitude, mais également des mesures sur des harmoniques. L'amplitude des harmoniques nous donne ainsi accès à la mesure $h_1 - h_2$ qui permet de calculer la pente spectrale et d'estimer le type de phonation du stimulus (voir partie 2.2.3).

Nous avons commencé par utiliser le mode *full-auto* du script : celui-ci détermine de lui-même toutes les mesures sans aucune intervention. Le paramétrage utilisé était : trois mesures par voyelle (une au début, une au milieu et une à la fin), pas de graphe en .pdf, l'option *iterate pulse*, et chercher dans 50ms à partir de chaque bordure. La Figure 17 représente l'interface du script.

Le constat que certains résultats obtenus étaient erronés (certaines mesures à -300 ou 0 pour les deux harmoniques par exemple) nous a poussé à passer le script en mode *manual* afin de vérifier

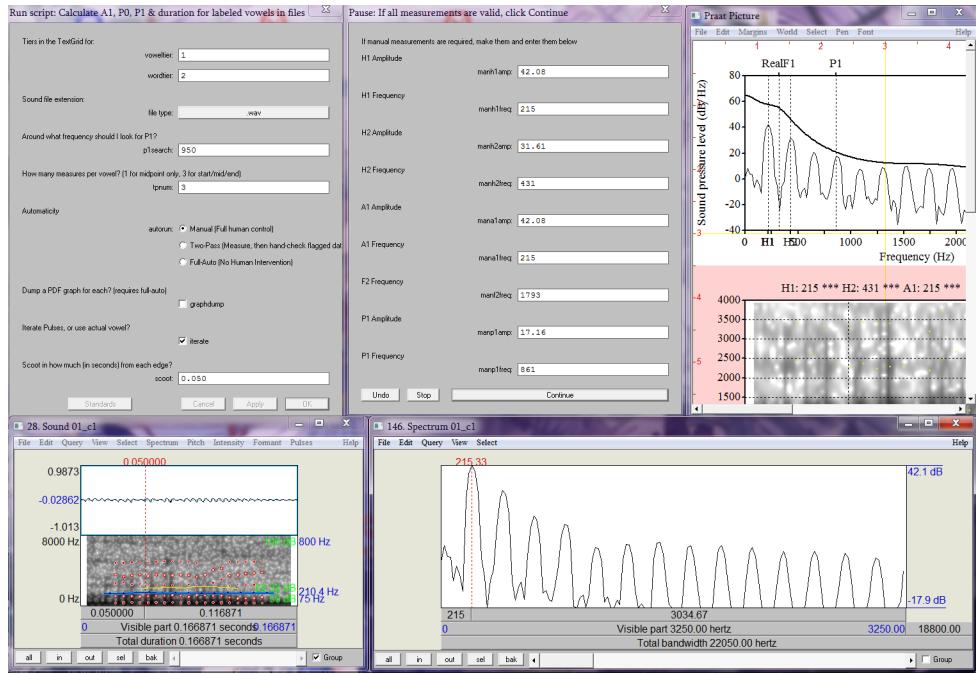


FIGURE 17 – Interface du script *NasalityAutomeasure* de Styler

chaque mesure prise par le script, et de comprendre d'où venaient les mesures erronées. Cette analyse est abordée dans la Partie acoustique des résultats.

Enfin, la troisième mesure provenait de notre propre script Praat (disponible https://github.com/C-Millot/memoire_m1/blob/main/detection_h1h2_manuelle.praat) pour des mesures manuelles obtenues différemment. Pour cela, on a sélectionné cinquante millisecondes au milieu de la voyelle (on a alors une seule mesure (qui sera la plus stable comparée à des mesures aux bords de la voyelle), contre trois pour le script automatique), dont on extrayait le spectre avec une fenêtre gaussienne. Ce fenêtrage permettait d'adoucir les bords de l'extraction afin de rendre le milieu de l'extrait plus important que les bords (voir Figure 18). En effet, les bords ont plus de chance d'être irréguliers que le milieu de la voyelle, et risqueraient de fausser les résultats. La particularité d'une fenêtre gaussienne est qu'elle adoucit les bords selon une courbe de Gauss, comme le montre la Figure 19.

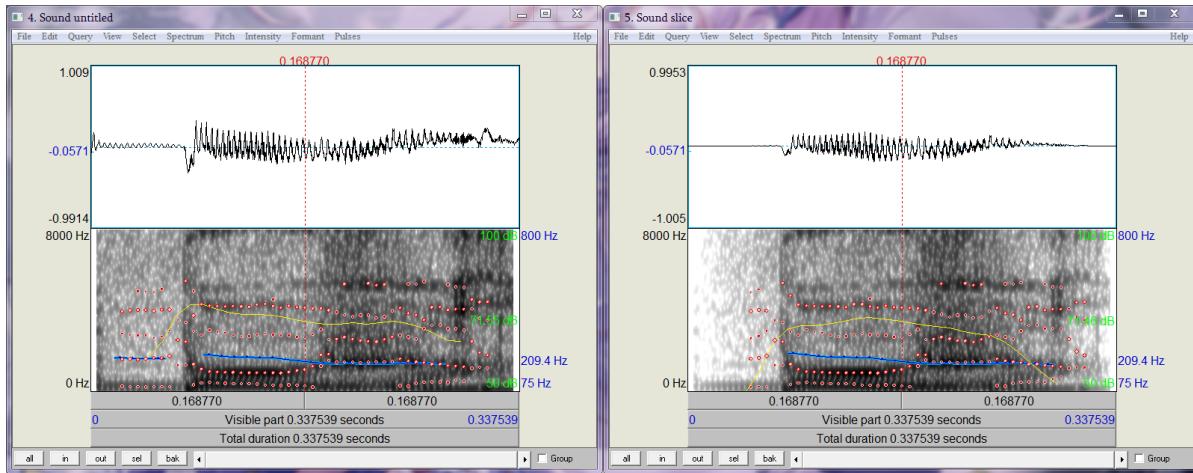


FIGURE 18 – Exemple d'un signal acoustique non modifié (à gauche) et fenêtré à l'aide d'une fenêtre gaussienne (à droite)

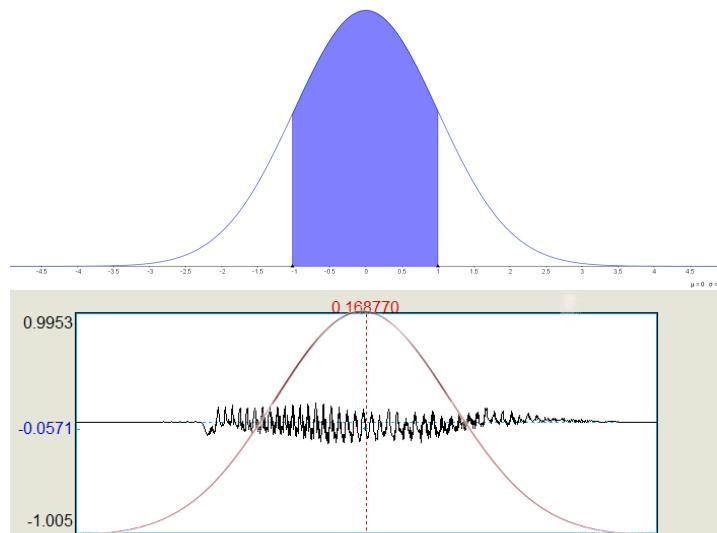


FIGURE 19 – Courbe de Gauss obtenue avec GeoGebra et visualisation sur un signal adouci à l'aide d'une fenêtre gaussienne

L'ouverture d'une fenêtre *demo* vierge permettait d'afficher le spectre, sur lequel l'utilisateur pouvait cliquer pour sélectionner la position du premier harmonique, puis du deuxième. Les coordonnées de cette position représentant la fréquence de l'harmonique (abscisses) et son amplitude (ordonnées), nous avons choisi d'enregistrer l'ordonnée de chaque harmonique sélectionné dans un fichier et de les soustraire. La Figure 20 représente l'interface du script.

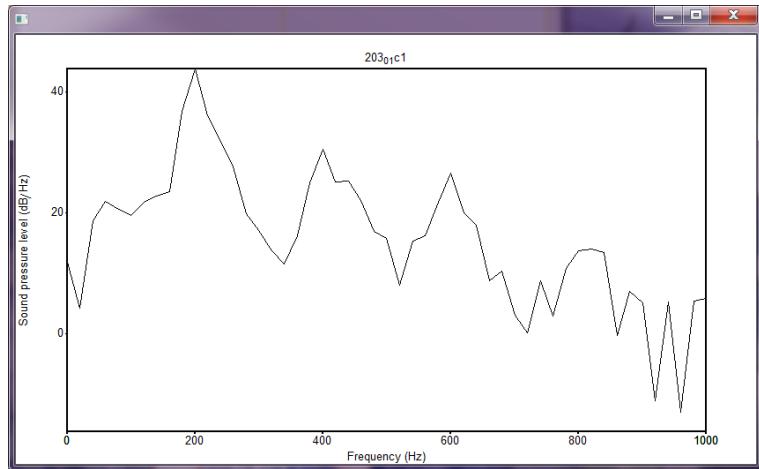


FIGURE 20 – Interface de notre script manuel

4.3.1 Résumé

Pour l'analyse acoustique des données, nous avons utilisé trois mesures :

- la version entièrement automatique du script Praat "NasalityAutomeasure" ;
- sa version manuelle ;
- notre propre script, entièrement manuel.

Cela nous permet de comparer les mesures entre elles, en plus de les comparer aux estimations des autres.

5 Résultats

Afin d'obtenir les résultats, une normalisation de ceux-ci était nécessaire : nous avons utilisé un script R rédigé par M.Audibert (https://github.com/C-Millot/memoire_m1/blob/main/traitement_reponses_test_psytoolkit_modifie.R) afin d'avoir tous les résultats PsyToolKit (participant, réponse au questionnaire, réponses au test) dans un même tableur.

Notre évaluation des résultats s'articulera autour des points suivants :

- l'accord entre les différents types de prédicteurs (participants au test, réseau de neurones et mesures acoustiques), et leurs performances selon le type de phonation, le locuteur, la voyelle utilisée...);
- les stimuli qui ont été particulièrement mal reconnus ;
- les particularités glottales éventuelles des différents locuteurs d'après les mesures.

5.1 Performances des mesures

Nous allons observer la performance des mesures acoustiques par rapport au type de phonation de chaque stimulus, tel qu'il a été annoté dans le corpus. Nous verrons aussi si l'un des scripts est plus performant, en termes de prédiction de type de phonation et selon le locuteur ou la voyelle produite.

5.1.1 Mesures acoustiques

Comme nous l'avions évoqué dans la Partie 4.3, pour certaines mesures nous avions des résultats surprenants. Par exemple, pour la version *full-auto* du script de Styler, nous avons remarqué que les deux stimuli en voix craquée du locuteur 8 avaient des 0 pour toutes les mesures (trois par voyelles). Pour d'autres, comme la troisième mesure du premier stimulus en voix craquée du locuteur 1, les amplitudes de h_1 et h_2 étaient toutes deux de -300, ce qui était étonnant.

Ces constats nous ont poussé à mettre le script en mode manuel, et nous avons pu voir ce qui lui posait problème. Comme on le voit sur la Figure 21, le script n'a détecté aucun signal sur l'extrait qu'il a sélectionné, et le spectre ne comporte donc aucun harmonique.

Ces problèmes n'arrivent presque que lors de l'analyse de voix craquée (et un peu lors de voix soufflée). Ces deux types de phonation — en particulier la voix craquée — sont connus pour leurs irrégularités dans le signal (voir Partie 2.2.2).

Le script se base sur une LPC (*Linear Predictive Coding*) afin d'obtenir le spectre : ce modèle

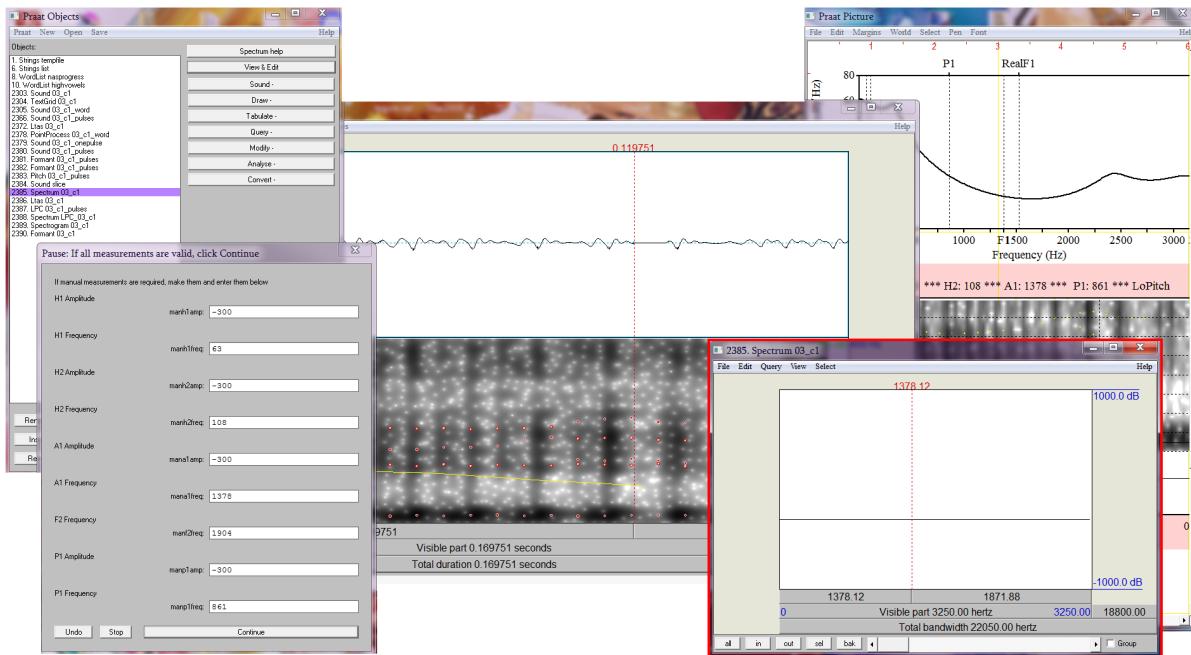


FIGURE 21 – Capture d'écran du script *NasalityAutoMeasure* lors de l'analyse du premier stimulus en voix craquée du locuteur 3

prédit qu'un échantillon donné peut être déterminé par une combinaison linéaire des échantillons qui le précédent, à une constante près. Il se base sur le modèle source-filtre (voir Partie 2.1.1) donc il permet d'obtenir les paramètres du filtre qui modifient le signal de la source selon cette théorie, c'est-à-dire les formants, anti-formants et les harmoniques entre autres. Cependant, de par le postulat de l'existence d'une source (glottale donc régulière) du modèle source-filtre, le fonctionnement d'une LPC peut être perturbé si la source est irrégulière. Cela explique que le script ait des difficultés quand il est confronté à de la voix craquée ou soufflée.

Notre script manuel n'a pas rencontré de problème dans l'extraction des mesures des stimuli. Néanmoins, il faut noter que le script *NasalityAutoMeasure* a avant tout été créé pour mesurer la nasalité, et pas particulièrement pour étudier les types de phonation comme la voix craquée. Les résultats pour chaque scripts sont trouvables ici : https://github.com/C-Millot/memoire_m1/blob/main/resultats_h1_h2_styler_auto.xlsx, https://github.com/C-Millot/memoire_m1/blob/main/resultats_h1_h2_styler_manuel.xlsx, https://github.com/C-Millot/memoire_m1/blob/main/resultats_h1_h2_manuel.xls.

Comme le script *NasalityAutoMeasure* faisait trois mesures par stimulus, nous avons fait la moyenne de ces mesures pour les comparer à la mesure unique de notre script. Voici le fichier de

comparaison entre les résultats des scripts : https://github.com/C-Millot/memoire_m1/blob/main/resultats_h1_h2_comparaison.xlsx.

Nous avons pu calculer la mesure acoustique *amplitude du premier harmonique - amplitude du second* sur la totalité des stimuli ainsi que par voyelle. Nous avons ensuite extrait des histogrammes de RStudio, grâce au script https://github.com/C-Millot/memoire_m1/blob/main/histogrammes.R, représentant la moyenne des mesures sur la totalité des stimuli (Figure 23), sur les deux voyelles étudiées (Figure 22).

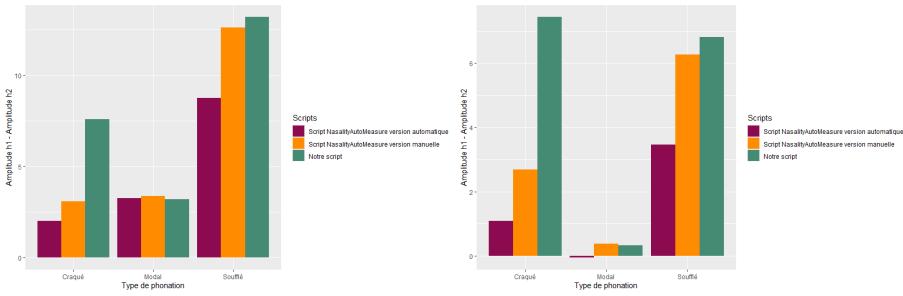


FIGURE 22 – Histogrammes groupés pour les voyelles /œ/ et /e/ représentant la mesure $h_1 - h_2$ pour les scripts *NasalityAutoMeasure* en version automatique, *NasalityAutoMeasure* en version manuelle, et notre propre *script manuel* pour chaque type de phonation (craqué, modal, soufflé)

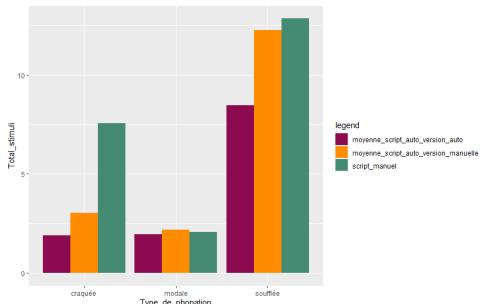


FIGURE 23 – Histogrammes groupés pour la totalité des stimuli représentant la mesure $h_1 - h_2$ pour les scripts *NasalityAutoMeasure* en version automatique, *NasalityAutoMeasure* en version manuelle, et notre propre *script manuel* pour chaque type de phonation (craqué, modal, soufflé)

Nous voyons que les prédictions ne sont pas très bonnes : d'après notre état de l'art Partie 2.2.3, un résultat négatif est craqué, un résultat positif est soufflé et un résultat proche de 0 est neutre. Sur la totalité des stimuli, la voix craquée a des résultats positifs (Figure 23). C'est la voix craquée qui est la plus sujette à de mauvaises prédictions — c'était prévisible, en effet nous avons vu que ce type de phonation posait des problèmes aux scripts de par son irrégularité. Toutefois, pour certains locuteurs comme les locuteurs 2, 3 (Figure 32), 4 et 5 (Figure 33) par exemple, la voix craquée est correctement identifiée par au moins un des scripts.

Si l'on différencie les résultats selon la voyelle (/e/ ou /œ/) (Figure 22), on voit qu'ils sont légèrement meilleurs pour /e/ surtout pour la voix modale.

Nous pouvons calculer le nombre et la moyenne de bonnes prédictions par script afin d'évaluer leurs performances globales. Pour cela, il faut déterminer un seuil au-delà duquel le type de phonation sera considéré comme mal prédit. Nous avons décidé que les résultats sous -2 seraient craqués, ceux au-dessus de 2 seraient soufflés, et ceux entre -2 et 2 seraient modaux. Le Tableau 2 récapitule les taux de bonnes prédictions selon chaque script et chaque voix. Les deux premiers scripts sont celui de Styler selon les différents paramètres utilisés.

	Script <i>full-auto</i>	Script auto manuel	Notre script
Nombre bonnes réponses Voix craquée	3/22	4/22	1/22
Moyenne $h_1 - h_2$ Voix craquée	1,81 > -2	2,895 > -2	7,51 > -2
Nombre bonnes réponses Voix modale	16/33	15/33	16/33
Moyenne $h_1 - h_2$ Voix modale	2 > 1,595 > -2	2 > 1,87 > -2	2 > 1,27 > -2
Nombre bonnes réponses Voix soufflée	17/18	18/18	18/18
Moyenne $h_1 - h_2$ Voix soufflé	6,11 > 2	9,435 > 2	10 > 2

TABLE 2 – Nombre de bonnes prédictions et moyenne de $h_1 - h_2$ selon chaque script et chaque voix. La couleur verte signifie que le script prédit en moyenne correctement le type de phonation ; le rouge que le script a en moyenne mal prédit le type de phonation

On voit que le type de phonation le mieux reconnu est la voix soufflée, et la voix modale, bien qu'elle soit souvent mal prédite, a un $h_1 - h_2$ moyen acceptable.

5.1.2 Résultats du test de perception

Les résultats du test de perception sont très bons ; seul un stimulus est mal deviné à plus de 50% (le deuxième stimulus soufflé du locuteur 11). Un participant moyen devine plus facilement la voix modale à presque 88%, suivie par la voix soufflée à environ 80% et la voix craquée à 67%.

Cependant, les performances des participants ne sont pas toutes égales. Par exemple, comme on peut le voir sur la Figure 24, le type de phonation le mieux reconnu par le sujet s.1277... est la voix soufflée, alors que s.5030... reconnaît bien mieux la voix modale que les deux autres. Aucun participant n'est plus performant pour la voix craquée que pour les autres types de phonation, mais cette reconnaissance peut tout de même aller jusqu'à 76% pour certains sujets, et ne les empêche pas forcément d'avoir des bons scores pour les autres types de phonation : s.489c... reconnaît aussi parfaitement la voix modale, et à 89% la voix soufflée. D'autres ont aussi de mauvais résultats de reconnaissance, comme le participant s.b68e... qui a eu moins de cinquante bonnes réponses sur

soixante-douze, contrairement aux autres qui ont au moins cinquante-sept. C'était aussi, et de loin, la personne avec les moins bons scores sur les échelles de Likert (aucune note au-dessus de "3" (moyennement sûr)), et on voit ainsi qu'elle était très peu sûre d'elle.

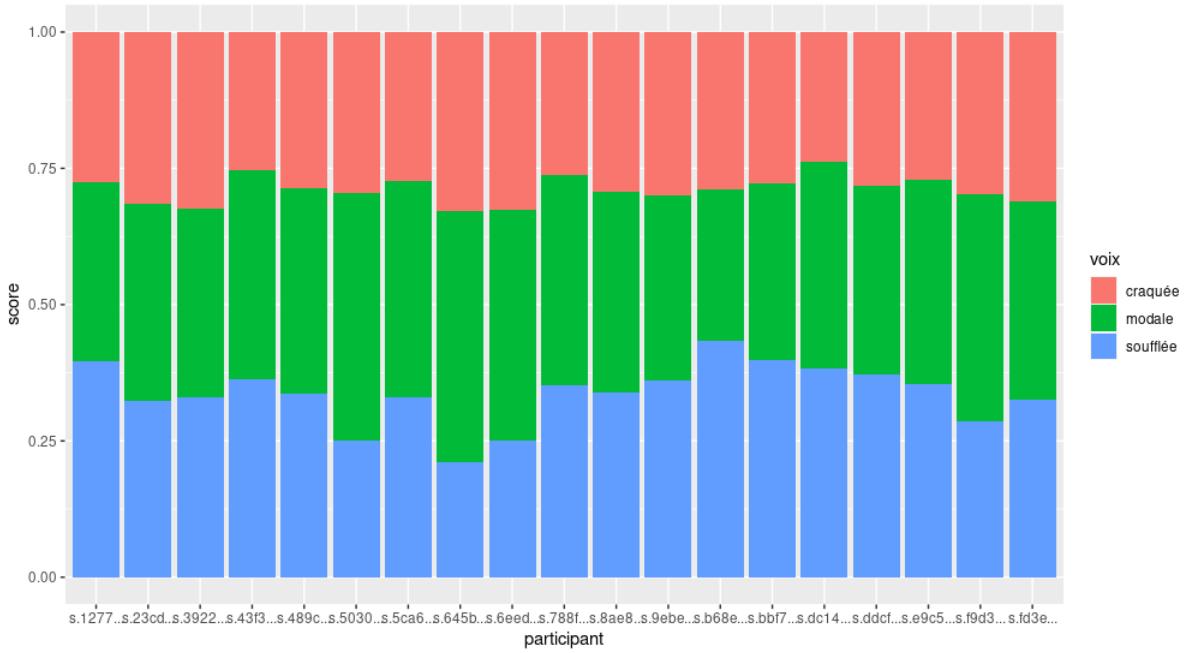


FIGURE 24 – Histogramme en barres empilées à pourcentage cumulé représentant le taux de bonnes réponses par participant selon le type de phonation

Cette utilisation des échelles de Likert n'était pas forcément la même pour tous les participants : certains étaient souvent très sûrs de leurs réponses, pour un score moyen par rapport aux autres participants au test. Il y avait deux attitudes qui se dégageaient des réponses aux échelles de Likert : la catégorie des personnes sûres d'elle attribuant souvent un "5" (meilleure note) à leurs réponses, et celle des personnes un peu moins sûres et réservées qui attribuaient souvent un "4". D'après quelques évaluations sur les données, il n'y a pas de corrélation entre le nombre de bonnes réponses au test et l'évaluation des réponses sur l'échelle de Likert.

5.1.3 Prédictions du réseau de neurones

Dans l'article de Chanclu et al. (2021), les auteurs font part d'un taux de bonnes prédictions d'environ 80% du réseau de neurones. Ce taux est plus haut pour les stimuli que nous avions spécifiquement sélectionnés : le réseau de neurones n'attribue que deux fois une prédition à moins de 60% de certitude pour le bon type de phonation (pour le deuxième stimulus en voix modale du locuteur 12, et pour le deuxième stimulus en voix soufflée du locuteur 8). De ce fait, la moyenne

de ses probabilités de réponse pour chaque type de phonation est de 95% pour la voix craquée, 96% pour la voix modale et 94% pour la voix soufflée.

C'est un taux de réussite très élevé, bien plus que celui du test de perception. Toutefois, en ce qui concerne les erreurs sous la barre des 50 à 60%, leur nombre est très similaire pour les deux prédicteurs, et le réseau de neurones a même davantage d'erreurs. Pour les deux prédicteurs, ces erreurs ne concernent jamais la voix craquée à l'inverse des mesures acoustiques.

Nous avons compilé les estimations du test de perception et du réseau de neurones dans des histogrammes groupés à l'aide de Rstudio, comme nous l'avions fait pour les mesures acoustiques.

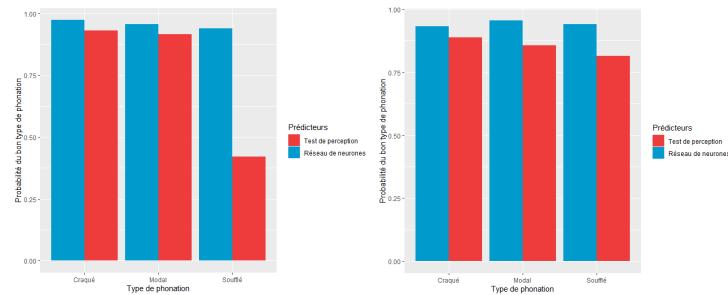


FIGURE 25 – Histogrammes groupés pour les voyelles /œ/ et /e/ représentant le taux de probabilité pour le bon type de phonation du [Réseau de neurones](#) et le pourcentage de bonnes réponses pour le [Test de perception](#) pour chaque type de phonation (craqué, modal, soufflé)

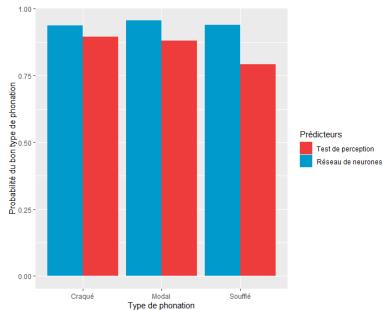


FIGURE 26 – Histogrammes groupés pour la totalité des stimuli représentant le taux de probabilité pour le bon type de phonation du [Réseau de neurones](#) et le pourcentage de bonnes réponses pour le [Test de perception](#) pour chaque type de phonation (craqué, modal, soufflé)

On voit que le réseau de neurones est très souvent au-dessus des participants au test de perception, que ce soit pour la totalité des stimuli (Figure 26) ou pour chaque voyelle séparément (Figure 25). Les participants ont très mal deviné la voix soufflée pour la voyelle /œ/ notamment, voyelle et type de phonation qui constituaient effectivement le stimulus le plus mal reconnu du test. Plus largement, ils ont moins bien deviné les stimuli soufflés. Le réseau de neurones a reconnu tous les

types de phonation plus ou moins également. Nous verrons dans la Partie 5.3 ce qu'il en est pour chaque locuteur.

5.1.4 Comparaison des performances

Le Kappa κ de Cohen est une mesure d'accord inter-annotateurs : elle nous permet de calculer à quel point deux prédicteurs ont annoté des stimuli similairement. De nombreuses interprétations existent concernant les résultats du Kappa de Cohen. D'après Landis and Koch (1977), un Kappa inférieur à 0 marque l'inexistence d'un accord, entre 0 et 0,2 il est faible, entre 0,2 et 0,4 il est passable, entre 0,4 et 0,6 il est modéré, entre 0,6 et 0,8 il est acceptable et au-dessus de 0,8 il est excellent.

Nous avons utilisé cet accord afin de comparer le taux de similarité entre les résultats de nos différentes mesures. Pour cela, nous avons ré-annoté nos données en remplaçant le type de phonation de chaque stimuli par celui qui était le plus reconnu, pour chaque type de mesure. Par exemple, si une voix craquée était davantage reconnue comme une voix soufflée, nous la renommions ainsi dans nos données.

	Test	CNN	Script <i>full-auto</i>	Script auto manuel	Notre script
Test	1	0,978	0,0414	-0,0558	-0,0495
CNN	0,978	1	0,0575	-0,0408	-0,0383
Script <i>full-auto</i>	0,0414	0,0575	1	0,643	0,408
Script auto manuel	-0,0558	-0,0408	0,643	1	0,537
Notre script	-0,0495	-0,0383	0,408	0,537	1

TABLE 3 – κ de Cohen effectué entre les prédictions du test de perception (Test), du réseau de neurones (CNN), des deux configurations du script Praat de Styler (Script *full-auto* et Script auto manuel) et de notre script Praat (Notre script)

D'après le Tableau 3, on voit que le test de perception et le réseau de neurones sont presque totalement d'accord entre eux. Cela n'est pas surprenant car nous avions vu qu'ils avaient deviné correctement presque tous les stimuli. Cependant, les résultats sont différents en ce qui concerne les accords inter-annotateurs impliquant les mesures acoustiques : les prédictions de chaque mesure acoustique comparées avec le test de perception ou le réseau de neurones sont très basses, proches de 0 et même parfois négatives. D'après Landis and Koch (1977), l'accord entre le test de perception ou le réseau de neurones et les mesures acoustiques est inexistant. Les deux modes du script *NasalityAutoMeasure* ont un bon accord, et leur Kappa de Cohen avec notre script est entre 0,4 et 0,6, ce qui est un accord modéré : ils ont donc des résultats différents, comme nous l'avions constaté

dans la Partie 2.2.3, mais néanmoins semblables.

5.1.5 Résumé

L'analyse des performances de chaque mesure nous révèle les points suivants :

- deux groupes se dessinent — les mesures acoustiques d'un côté et les participants humains et le réseau de neurones de l'autre ;
- ce deuxième groupe prédit bien mieux le type de phonation des stimuli ;
- les scripts peinent à obtenir des mesures cohérentes pour la voix craquée.

5.2 Analyse fine des stimuli

Il est intéressant d'aller effectuer des analyses fines sur quelques résultats attirant l'œil. Par exemple, les stimuli particulièrement mal reconnus. Nous avions déjà vu que, pour les mesures acoustiques, cela était toujours dû aux irrégularités dans la voix, souvent à cause de voix craquée. Étudions maintenant les stimuli mal reconnus par les participants au test de perception, et par le réseau de neurones.

Pour certains stimuli, les baisses de performance des participants sont en adéquation avec les mesures acoustiques : par exemple, le deuxième stimulus modal du locuteur 12, reconnu à seulement 52%, est très souvent confondu avec une voix soufflée ; les mesures acoustiques témoignent en effet d'un $h_1 - h_2$ particulièrement haut (jusqu'à 12), qui pourraient donc être corrélées à ce problème de perception. De plus, le réseau de neurones a également une performance moins bonne pour ce stimulus, et hésite aussi avec une voix soufflée.

À l'inverse, le stimulus en voix modale du locuteur 17 est confondu par les participants avec une voix craquée, et les mesures sont justement négatives pour celui-ci. Cependant, le réseau de neurones n'est cette fois pas en accord avec ces estimations et prédit correctement à 98% la voix modale (et la voix craquée à seulement 1%). Enfin, un dernier exemple de prédictions humaines erronées congrues aux mesures acoustiques est le stimulus soufflé du locuteur 5 : les participants du test le confondent à ~40% avec une voix modale, et les mesures sont effectivement proches de 0. Le réseau de neurones, lui, reconnaît la voix soufflée à 89% et n'accorde que 8% à la voix modale.

Pour d'autres extraits, les mesures acoustiques n'expliquent toutefois pas les difficultés rencontrées par les participants : c'est le cas du stimulus soufflé du locuteur 8 par exemple, ou encore

du deuxième du locuteur 11... Nous étudierons leur cas lors de la Discussion.

5.3 Évaluation des types de phonation des locuteurs

Dans cette partie, nous allons voir si les différentes mesures de prédiction que nous avons utilisées (test de perception, mesures acoustiques, réseau de neurones) nous permettent d'identifier un type de phonation particulièrement saillant selon le locuteur.

Pour cela nous avons à la manière des parties précédentes, construit des histogrammes groupés, pour les trois mesures acoustiques en fonction de $h_1 - h_2$ (exemple avec la Figure 27) et pour nos deux mesures plus proches de la perception en fonction de la probabilité du bon type de phonation prédite (exemple avec la Figure 28). Les autres figures sont disponibles en Annexe.

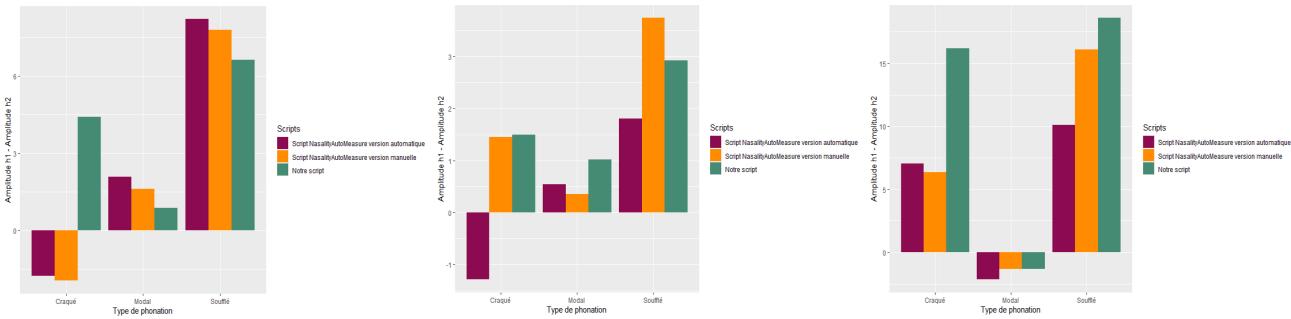


FIGURE 27 – Histogrammes groupés pour les locuteurs 4, 5 et 6 représentant la mesure $h_1 - h_2$ pour les scripts *NasalityAutoMeasure* en version automatique, *NasalityAutoMeasure* en version manuelle, et notre propre script manuel pour chaque type de phonation (craqué, modal, soufflé)

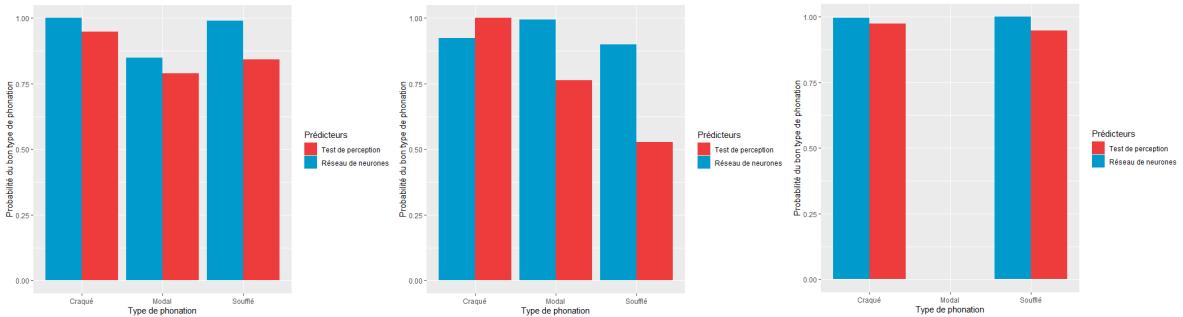


FIGURE 28 – Histogrammes groupés pour les locuteurs 4, 5 et 6 représentant le taux de probabilité pour le bon type de phonation du Réseau de neurones et le pourcentage de bonnes réponses pour le Test de perception pour chaque type de phonation (craqué, modal, soufflé)

Concernant les mesures acoustiques, pour le locuteur 1 (Figure 32) les trois mesures sont bien au-dessus de 0 pour tous les types de phonation. Pour plusieurs locuteurs comme les locuteurs 3 (Figure 32) et 4 (Figure 33), notre script manuel peine à identifier leur voix craquée et fournit des

mesures bien au-dessus de 0 contrairement à celui de Styler. Les mesures modales sont souvent plus basses que celles soufflées, ce qui est congru à ce qui est attendu normalement.

On voit que les scripts sont meilleurs pour certains locuteurs : notamment, le locuteur 2 (Figure 32) et le 3 pour *NasalityAutoMeasure* (Figure 32). Pour certains autres, on voit que les mesures sont toujours particulièrement élevées : c'est le cas pour les locuteurs 1 (Figure 32), 13 (Figure 36) et 16 (Figure 37) par exemple. Et pour d'autres, enfin, les mesures sont très négatives : locuteur 17 (Figure 37) et 20 (Figure 38).

Le réseau de neurones comme les participants humains reconnaissent bien la plupart des stimuli des locuteurs. On pourrait penser que des mesures acoustiques négatives traduiraient une pente spectrale moins importante, et donc une voix craquée proéminente chez le sujet qui serait mieux perçue par des auditeurs ; et inversement pour la voix soufflée. Qu'en est-il ici ?

Pour le locuteur 17 (Figure 45), dont les mesures acoustiques pour la voix modale sont très basses, on voit en effet que les participants sont perturbés et devinent mal le type de phonation. Pour le locuteur 20 (Figure 46) qui a ce même type de mesures, la voix craquée est mieux reconnue que la moyenne par les participants. Toutefois, la voix modale est bien reconnue aussi, donc il est possible que ce ne soit que dû au hasard. Dans tous les cas, il y a bien certains locuteurs dont la voix craquée est bien reconnue, et les autres types de phonation le sont bien moins : c'est le cas des locuteurs 4, 5, 8, 10 et 11 notamment. Peut-être que cela peut indiquer une voix craquée proéminente, au point de complexifier l'identification des autres types de phonation.

Cela peut se vérifier en allant voir le détail des confusions des participants ; il s'avère que les locuteurs 4 et 5 n'ont pas leurs stimuli particulièrement confondus avec de la voix craquée, cependant c'est notamment le cas du locuteur 8, ce qui explique peut-être le résultat que nous avions évoqué avant en ce qui concerne son stimulus en voix soufflée — il a peut-être une voix particulièrement craquée. Les cas des locuteurs 10 et 11 sont mitigés.

Le locuteur 1 (Figure 40), de façon intéressante, a sa voix soufflée très bien reconnue par rapport aux autres locuteurs, et surtout elle est mieux reconnue que les autres types de phonation, ce qui est rarement le cas. Peut-être que ce locuteur a une qualité de voix plus soufflée que les autres locuteurs. Cela se retrouve dans les mesures acoustiques, très élevées pour tous les types de phonation sans exception, et bien plus que pour tous les autres locuteurs. La seule autre fois où ces mesures se reproduisent est pour le locuteur 16 (Figure 45), où les participants adoptent la même attitude

que pour le locuteur 1, et reconnaissent mieux sa voix soufflée. D'après le détail des réponses des participants, la plupart des confusions sont dues à la mauvaise prédiction de la voix soufflée pour les deux autres types de phonation.

Le réseau de neurones n'est pas autant influencé par ces phénomènes que les auditeurs humains.

5.3.1 Résumé

Toue les locuteurs n'ont pas de type de phonation plus présent qu'un autre, mais cela peut arriver :

- les mesures acoustiques ont trop de problèmes avec la voix craquée pour apporter de l'aide dans la compréhension de la perception de ce type de phonation ;
- mais dans le cas de mesures très élevées, celles-ci peuvent indiquer, tout comme la perception des locuteurs, la présence d'un locuteur à la voix plus soufflée que d'ordinaire, plus reconnaissable ;
- une très bonne reconnaissance d'un type de phonation en moyenne ne veut pas dire que la production vocale d'un locuteur est influencée par celui-ci (qualité de voix particulière) et que les autres types de phonation seront mal reconnus par conséquent. Pour savoir si c'est le cas, il faut aller voir en détail la confusion des locuteurs ;
- certains locuteurs ont leur qualité de voix particulièrement influencée par un type de phonation d'après la perception de certains locuteurs.

6 Discussion

Nous avons pu voir dans les résultats une grande disparité entre les évaluations du test de perception et du réseau de neurones, et les mesures acoustiques.

Il est intéressant de noter que pour les évaluations par les participants au test de perception et par le réseau de neurones, les voix craquées sont les mieux reconnues et aucune confusion à plus de 50% n'est observée. Inversement pour les mesures acoustiques, les voix craquées sont toujours celles qui sont le moins bien prédites. La teneur très irrégulière de la voix craquée est peut-être plus appréciable en prenant en compte tout l'extrait, plutôt que quelques points d'intérêt comme le fonctionnement des mesures acoustiques. Au sujet du test de perception, il faut noter que les participants les plus confiants en leurs réponses n'étaient pas forcément ceux qui répondaient le mieux : il faut donc rester méfiant de la façon dont chacun s'auto-évalue.

La mesure de l'amplitude de h_2 soustraite à l'amplitude de h_1 n'est pas adéquate pour déterminer le type de phonation d'un extrait sonore, particulièrement pour la voix craquée. Toutefois, ce n'est pas que la mesure en elle-même qui émet de mauvaises prédictions, mais surtout la façon dont les harmoniques sont obtenus — à l'aide d'une LPC pour approximer la f_0 . En trouvant une façon plus stable de calculer h_1 et h_2 , les résultats seraient peut-être plus probants, et nos observations ne peuvent ainsi pas complètement remettre en doute les analyses de Keating et al. (2010) par exemple. Le réseau de neurones offre cependant des prédictions bien plus justes que les scripts Praat, et même les participants au test de perception — rappelons cependant que les participants sont parfois influencés par une possible qualité de voix du locuteur.

En plus de ces différentes qualités de voix provoquant de la variation inter-locuteurs, il existe aussi des variations intra-locuteurs, principalement remarquables lorsque l'on compare le taux de reconnaissance de deux stimuli du même type de phonation produits par le même locuteur. Par exemple, sur les deux stimuli craqués du locuteur 24, le premier est mieux reconnu que le deuxième, que ce soit pour les participants au test comme le réseau de neurones. Cela est aussi le cas des deux stimuli modaux du locuteur 12, dont le deuxième est particulièrement mal reconnu par les deux prédicteurs. Dans les deux cas, les stimuli dans chaque paire ont des durées et des voyelles semblables et les locuteurs ne pouvaient pas être influencés par ces facteurs.

Enfin, en complément de l'analyse fine de certains stimuli, nous souhaitons aller observer ceux

dont la mauvaise reconnaissance n'a pas pu être expliquée par les mesures acoustiques.

Nous avons recherché la production soufflée du locuteur 8 dans l'enregistrement original afin de la remettre en contexte.

En réécoutant le /œ/ seul, nous percevons un aspect soufflé mais, à la toute fin, un aspect craqué également. Effectivement, lorsque l'on écoute la suite de l'enregistrement, nous nous apercevons que la production qui suit la voyelle final est craquée (comme on peut le voir sur le spectrogramme d'une partie de l'enregistrement, Figure 29) ; c'est le début de la voix craquée qui suit la voyelle que nous entendons et qui a influencé tant de participants ; cela montre que l'évolution du type de phonation dans la voyelle est perceptible et utilisé par les auditeurs dans des tâches de différentiation. Cela explique aussi que les mesures ne fonctionnent pas sur le stimulus : elles se servent de moyennes calculées pour effectuer des prédictions, or ici le changement est important dans la bonne perception.

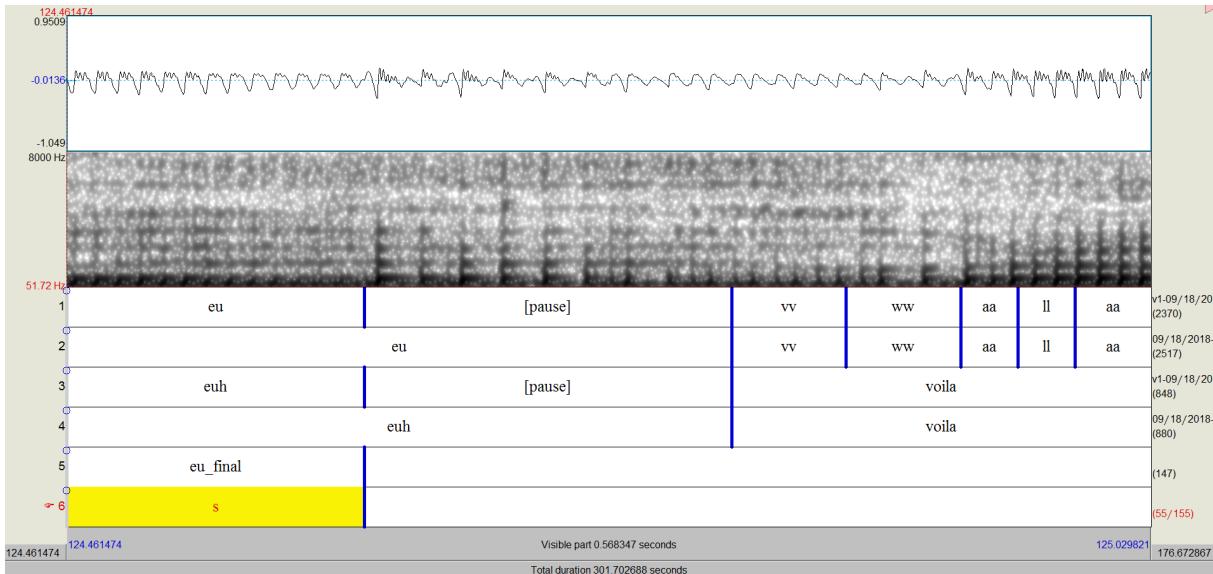


FIGURE 29 – Spectrogramme montrant le contexte du stimulus soufflé du locuteur 8

Le spectrogramme commence déjà à être un peu craqué durant le stimulus, peut-être les auditeurs s'en sont-ils donc servi afin d'identifier le type de phonation. L'annotation du corpus aurait dû être craquée plutôt que soufflé, étant donné le contexte d'énonciation.

Pour le deuxième stimulus soufflé du locuteur 11, voici ce que nous avons trouvé :

Nous voyons sur la Figure 30 que la délimitation du stimulus a été faîte avant la fin de celui-ci, et sa qualité soufflée en a sans doute pâti pour cette raison. À titre de comparaison, voici un stimulus soufflé produit par le locuteur 16, parfaitement reconnu par tous les locuteurs, les mesures acoustiques et le réseau de neurones :

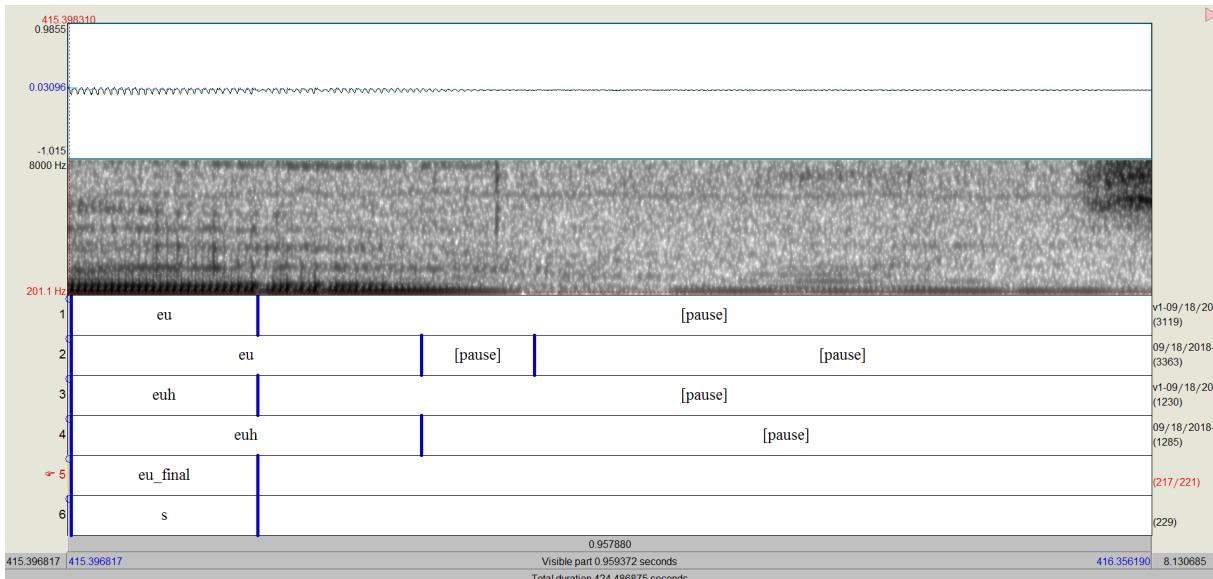


FIGURE 30 – Spectrogramme montrant le contexte du stimulus soufflé du locuteur 8

On distingue parfaitement la composante soufflée de la production vocale sur ce spectrogramme Figure 31.

Pour un des deux stimuli mal reconnus que nous avons observés, et pour peut-être d'autres, la cause de la reconnaissance faible est que le type de phonation cible n'était peut-être pas le bon ; on peut alors penser à une erreur d'annotation, ce qui est compréhensible tant la tâche peut se montrer longue et répétitive. L'autre stimuli a subi une autre sorte d'erreur d'annotation, en étant coupé avant sa fin sans pouvoir révéler sa véritable composante soufflée.

6.1 Critiques

Certaines critiques peuvent être formulées au sujet de notre recherche. D'abord, en ce qui concerne le test de perception, celui aurait pu être parfait à bien des égards : les stimuli, par exemple, n'auraient pas dû être présentés bruts aux oreilles des auditeurs, et auraient pu bénéficier d'un fondu à leurs début et fin, éventuellement d'un bruit blanc avant afin d'avertir le participant du commencement du stimulus, et d'une égalisation de l'intensité sur tout notre corpus. Nous espérons que ces problèmes ont été atténués par la possibilité de ré-écouter le stimulus au besoin. Ayant recueilli les impressions de certains participants, nous avons également constaté certains défauts de fluidité, de moments de flottement lors du test, et aussi parfois de la mauvaise compréhension de la consigne (« Avait-on le droit de lire les spectrogrammes ? »), défauts que nous pourrons améliorer

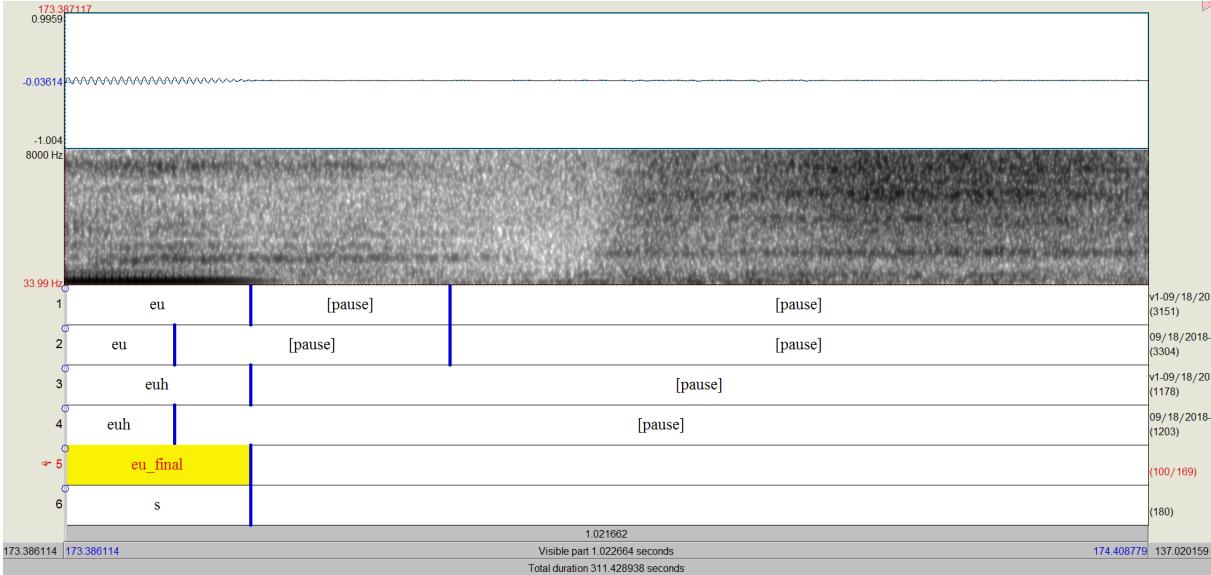


FIGURE 31 – Spectrogramme montrant le contexte du stimulus soufflé du locuteur 8

lors d'une prochaine construction de test. Peut-être qu'un plus grand nombre de pré-testeurs serait intéressant. Enfin, une mesure du temps de réaction des participants nous aurait intéressé a posteriori, car elle aurait pu nous montrer si un type de phonation particulier venait plus ou moins vite à l'esprit des auditeurs. Elle rentrait toutefois en conflit avec la consigne d'écouter le stimulus autant de fois que nécessaire sans se presser (consigne établie notamment à cause du nombre important de stimuli de l'étude, possible de fatiguer les participants).

Concernant les mesures acoustiques, il pourrait être intéressant de tester d'autres mesures comme la prominence cepstrale. Nous n'avons utilisé que $h_1 - h_2$ en raison de sa popularité et qu'elle est réputée pour ses bons résultats.

Notre corpus était adapté à notre étude, toutefois la présence de mesures articulatoires comme les périodes d'ouverture des plis vocaux mesurées par un électroglottographe aurait pu présenter l'avantage d'étayer nos hypothèses sur des qualités de voix différentes selon les locuteurs et parfois perceptibles par les auditeurs.

L'inconvénient de ces mesures est qu'il est moins facile de produire une parole tout à fait spontanée et libre.

7 Conclusion

Notre étude s'articulait autour des deux hypothèses suivantes :

- 1) Certains locuteurs ont un type de phonation plus saillant et reconnaissable que d'autres.
- 2) Les différentes méthodes de classification des types de phonation (humain, $h_1 - h_2$, réseau de neurones) ne sont pas toutes aussi performantes, et on s'attend à de meilleures performances d'un réseau de neurones.

D'après nos résultats, la première hypothèse ne peut pas être invalidée. En effet, des locuteurs comme le locuteur 1 (voix soufflée) ou 8 (voix craquée) provoquent des confusions de la part des auditeurs pour les autres types de phonation, et ont un de ces types particulièrement bien reconnu.

Concernant la deuxième hypothèse, nous avons vu que les mesures avaient des résultats très différents, avec le réseau de neurones comme le plus performant. Cependant, malgré la littérature sur le sujet, les auditeurs humains ont eu des performances bien meilleures que les mesures acoustiques prises à l'aide de scripts Praat, notamment à cause de difficultés techniques liées à la f_0 , mais aussi car la voix craquée est un type de phonation que les meilleurs prédicteurs apprécient davantage sur toute la longueur de l'extrait et pas seulement au niveau de points d'intérêt.

Afin de clarifier nos résultats, une suite à cette étude pourrait constituer en l'obtention de mesures articulatoires de la part des locuteurs, nous permettant de savoir si les résultats observés peuvent réellement être liés à des particularités physiques ou non.

De plus, il serait intéressant de voir si la distinction d'un type de phonation affectant la qualité de voix d'un locuteur, permet de le reconnaître et de s'en souvenir plus facilement. Pour cela, on pourrait construire un test de perception à l'image d'un de ceux présentés dans Greenberg et al. (2010) : il s'agirait de mettre deux courts extraits l'un à la suite de l'autre, et de demander aux participants s'ils étaient produits par le même locuteur. On s'attendrait alors à davantage de réponses "Oui" pour les locuteurs 1, 8, ou 16 par exemple, et cela pourrait varier en fonction du type de phonation le plus proéminent du locuteur, mais aussi d'autres aspects de sa qualité de voix qu'il faudrait contrôler (voix plus nasale par exemple).

8 Bibliographie

- Anderson, R. C., Klofstad, C. A., Mayew, W. J., and Venkatachalam, M. (2014). Vocal fry may undermine the success of young women in the labor market. *PloS one*, 9(5) :e97506.
- Andruski, J. E. and Ratliff, M. (2000). Phonation types in production of phonological tone : The case of green mong. *Journal of the International Phonetic Association*, 30(1-2) :37–61.
- Aristote (-324 avJ.C.). *Les Politiques*. FLAMMARION (April 22, 2015).
- ATILF - CNRS & Université de Lorraine (s.d.). TlfI : Trésor de la langue française informatisé. <http://www.atilf.fr/tlfI> [consulté le 23/06/2022].
- Baird, B. O. (2011). Phonetic and phonological realizations of 'broken glottal' vowels in k'ichee'. *Proceedings of formal approaches to Mayan linguistics : MIT working papers in linguistics*, 63 :39–49.
- Barsties, B. and De Bodt, M. (2015). Assessment of voice quality : current state-of-the-art. *Auris Nasus Larynx*, 42(3) :183–188.
- Benoist-Lucy, A. and Pillot-Loiseau, C. (2013). The influence of language and speech task upon creaky voice use among six young american women learning french. In *Interspeech 2013*, pages 2395–2399. International Speech Communication Association.
- Chanclu, A., Amor, I. B., Gendrot, C., Ferragne, E., and Bonastre, J.-F. (2021). Automatic classification of phonation types in spontaneous speech : towards a new workflow for the characterization of speakers' voice quality. In *Interspeech 2021*, pages 1015–1018. ISCA.
- Chanclu, A., Georgeton, L., Fredouille, C., and Bonastre, J.-F. (2020). Ptsvox : une base de données pour la comparaison de voix dans le cadre judiciaire. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*, pages 73–81. ATALA ; AFCP.
- Chernyak, B. R., Simon, T. B., Segal, Y., Steffman, J., Chodroff, E., Cole, J. S., and Keshet, J. (2022). Deepfry : Identifying vocal fry using deep neural networks. *arXiv preprint arXiv :2203.17019*.

- Denes, P. B., Denes, P., and Pinson, E. (1993). *The speech chain*. Macmillan.
- Drugman, T., Kane, J., and Gobl, C. (2020). Data-driven detection and analysis of the patterns of creaky voice. *arXiv preprint arXiv :2006.00518*.
- Epstein, M. A. (2002). *Voice quality and prosody in English*. University of California, Los Angeles.
- Esling, J. H. (1984). Laryngographic study of phonation type and laryngeal configuration. *Journal of the International Phonetic Association*, 14(2) :56–73.
- Esposito, C. M. (2005). An acoustic and electroglottographic study of phonation in santa ana del valle zapotec. In *Poster presented at the 79th meeting of the Linguistic Society of America, San Francisco, CA*.
- Esposito, C. M. and Khan, S. u. D. (2020). The cross-linguistic patterns of phonation types. *Language and Linguistics Compass*, 14(12) :e12392.
- Fant, G. (1981). The source filter concept in voice production. *STL-QPSR*, 1(1981) :21–37.
- Gerlach, L., McDougall, K., Kelly, F., Alexander, A., and Nolan, F. (2020). Exploring the relationship between voice similarity estimates by listeners and by an automatic speaker recognition system incorporating phonetic features. *Speech Communication*, 124 :85–95.
- Gibson, T. A. (2017). The role of lexical stress on the use of vocal fry in young adult female speakers. *Journal of Voice*, 31(1) :62–66.
- Gordon, M. and Ladefoged, P. (2001). Phonation types : a cross-linguistic overview. *Journal of phonetics*, 29(4) :383–406.
- Greenberg, C. S., Martin, A. F., Brandschain, L., Campbell, J. P., Cieri, C., Doddington, G. R., and Godfrey, J. J. (2010). Human assisted speaker recognition in nist sre10. In *Odyssey*, page 32.
- Greer, S. D. and Winters, S. J. (2015). The perception of coolness : Differences in evaluating voice quality in male and female speakers. In *ICPhS*.
- Hanson, H. and Chuang, E. (1999). Glottal characteristics of male speakers : Acoustic correlates and comparison with female data. *The Journal of the Acoustical Society of America*, 106 :1064–77.
- Hay, J. and Drager, K. (2007). Sociophonetics. *Annu. Rev. Anthropol.*, 36 :89–103.

- Hay, J. and MacLagan, M. (2006). Are all /r/s alike? degrees of constriction in new zealand english intrusive /r/. In *NZ Lang. Soc. Conf., Christchurch*.
- Holmberg, E. B., Hillman, R. E., Perkell, J. S., Guiod, P. C., and Goldman, S. L. (1995). Comparisons among aerodynamic, electroglossographic, and acoustic spectral measures of female voice. *Journal of Speech, Language, and Hearing Research*, 38(6) :1212–1223.
- Ishi, C. T., Sakakibara, K.-I., Ishiguro, H., and Hagita, N. (2007). A method for automatic detection of vocal fry. *IEEE transactions on audio, speech, and language processing*, 16(1) :47–56.
- Jessen, M. (2008). Forensic phonetics. *Language and linguistics compass*, 2(4) :671–711.
- Jiang, J., Lin, E., and Hanson, D. G. (2000). Vocal fold physiology. *Otolaryngologic Clinics of North America*, 33(4) :699–718.
- Kane, J. and Gobl, C. (2011). Identifying regions of non-modal phonation using features of the wavelet transform. In *Twelfth Annual Conference of the International Speech Communication Association*.
- Keating, P., Esposito, C., Garellek, M., Khan, S., and Kuang, J. (2010). Phonation contrasts across languages. In *Poster presented at the 12th Conference on Laboratory Phonology*.
- Kikura, M., Suzuki, K., Itagaki, T., Takada, T., and Sato, S. (2007). Age and comorbidity as risk factors for vocal cord paralysis associated with tracheal intubation. *British journal of anaesthesia*, 98(4) :524–530.
- Kreiman, J., Vanlancker-Sidtis, D., and Gerratt, B. R. (2003). Defining and measuring voice quality. In *ISCA Tutorial and Research Workshop on Voice Quality : Functions, Analysis and Synthesis*.
- Künzel, H. J. (2007). Non-contemporary speech samples : Auditory detectability of an 11 year delay and its effect on automatic speaker identification. *International Journal of Speech, Language & the Law*, 14(1).
- Ladefoged, P. (1971). Preliminaries to linguistic phonetics. *Chicago : University of Chicago*.
- Ladefoged, P. (1983). The linguistic use of different phonation types. *Vocal fold physiology : Contemporary research and clinical issues*, 351360.
- Ladefoged, P. and Maddieson, I. (1996). *The Sounds of the World's Languages*. Blackwell Publishers.

- Lancer, J. M., Syder, D., Jones, A., and Le Boutillier, A. (1988). Vocal cord nodules : a review. *Clinical Otolaryngology & Allied Sciences*, 13(1) :43–51.
- Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.
- Laver, J. (1980). The phonetic description of voice quality. *Cambridge Studies in Linguistics London*, 31 :1–186.
- Maddieson, I. (2007). Peter ladefoged.
- Mendoza-Denton, N. (2011). The semiotic hitchhiker's guide to creaky voice : Circulation and gendered hardcore in a chicana/o gang persona. *Journal of Linguistic Anthropology*, 21(2) :261–280.
- Morrison, G. S. (2009). Forensic voice comparison and the paradigm shift. *Science & Justice*, 49(4) :298–308.
- Podesva, R. J. (2006). Phonetic detail in sociolinguistic variation : Its linguistic significance and role in the construction of social meaning. palo alto, ca : Stanford university ph. d. *Journal of Sociolinguistics*, 1(1) :4.
- Podesva, R. J. and Callier, P. (2015). Voice quality and identity. *Annual review of applied Linguistics*, 35 :173–194.
- Pullens, B., Hakkesteegt, M., Hoeve, H., Timmerman, M., and Joosten, K. (2017). Voice outcome and voice-related quality of life after surgery for pediatric laryngotracheal stenosis. *The Laryngoscope*, 127(7) :1707–1711.
- Shue, Y.-L., Keating, P., Vicenik, C., and Yu, K. (2010). Voicesauce : A program for voice analysis. *Energy*, 1(H2) :H1–A1.
- Smith, C. L. (2002). Prosodic finality and sentence type in french. *Language and Speech*, 45(2) :141–178.
- Starr, R. and Greene, R. (2006). Beyond cuteness : The role of voice quality in performing stylized femininities in japanese. *New Ways of Analyzing Variation*, 36.
- Stevens, K. N. (1977). Physics of laryngeal behavior and larynx modes. *Phonetica*, 34(4) :264–279.

- Stoet, G. (2010). Psytoolkit - a software package for programming psychological experiments using linux. *Behavior Research Methods*, 42(4) :1096–1104.
- Stoet, G. (2017). Psytoolkit : A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1) :24–31.
- Tarafder, K. H., Datta, P. G., and Tariq, A. (2012). The aging voice. *Bangabandhu Sheikh Mujib Medical University Journal*, 5(1) :83–86.
- Vishnubhotla, S. and Espy-Wilson, C. Y. (2006). Automatic detection of irregular phonation in continuous speech. In *Ninth International Conference on Spoken Language Processing*.
- Watt, D. and Llamas, C. (2010). The identification of the individual through speech. *Language and identities*, pages 76–85.
- White, H., Penney, J., Gibson, A., Szakay, A., and Cox, F. (2021). Optimizing an automatic creaky voice detection method for australian english speaking females. In *Interspeech*, pages 1384–1388.
- Wright, R., Manfield, C., and Panfili, L. (2019). Voice quality types and uses in north american english. *Anglophonia [Online]*, 27.
- Yuasa, I. P. (2010). Creaky voice : A new feminine voice quality for young urban-oriented upwardly mobile american women ? *American Speech*, 85(3) :315–337.

9 Annexe

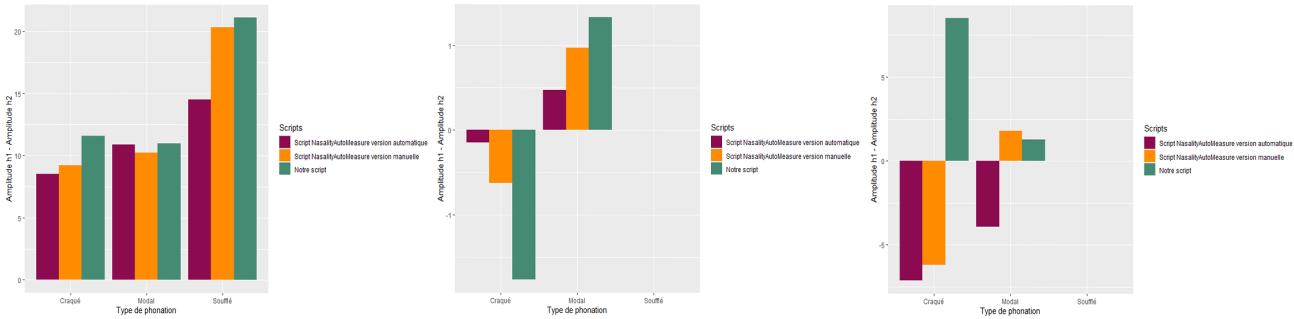


FIGURE 32 – Histogrammes groupés pour les locuteurs 1, 2 et 3 représentant la mesure $h_1 - h_2$ pour les scripts *NasalityAutoMeasure* en version automatique, *NasalityAutoMeasure* en version manuelle, et notre propre *script* manuel pour chaque type de phonation (craqué, modal, soufflé)

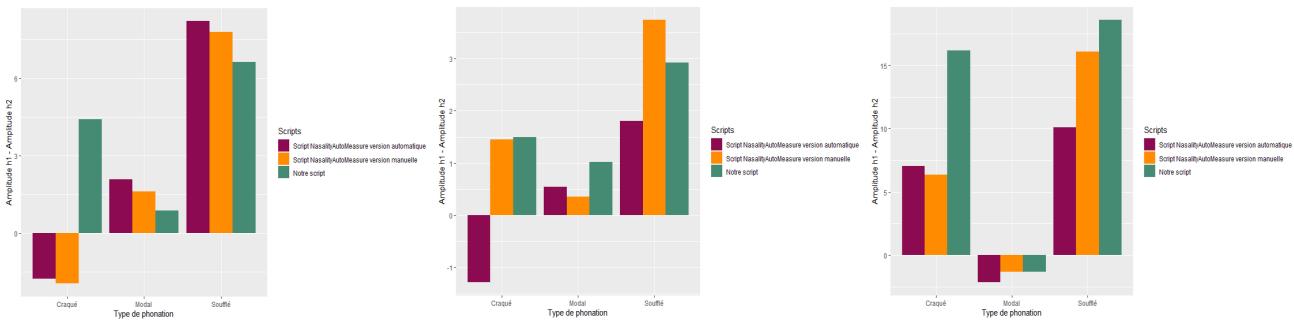


FIGURE 33 – Histogrammes groupés pour les locuteurs 4, 5 et 6 représentant la mesure $h_1 - h_2$ pour les scripts *NasalityAutoMeasure* en version automatique, *NasalityAutoMeasure* en version manuelle, et notre propre *script* manuel pour chaque type de phonation (craqué, modal, soufflé)

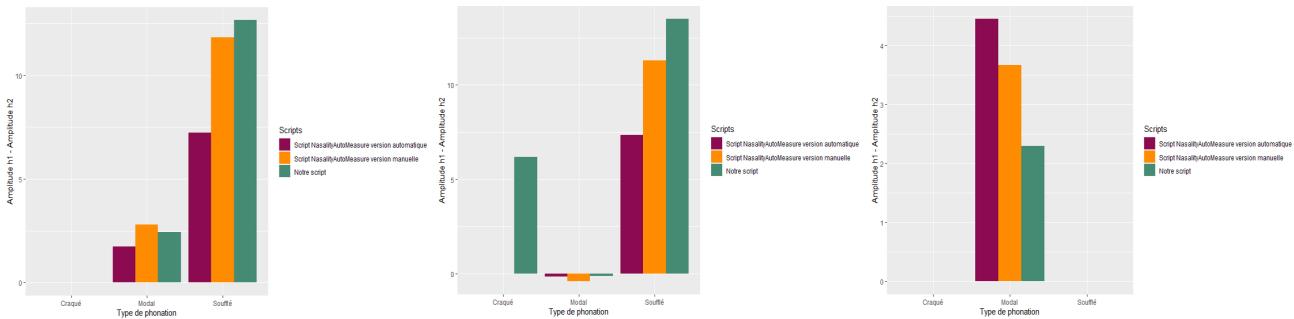


FIGURE 34 – Histogrammes groupés pour les locuteurs 7, 8 et 9 représentant la mesure $h_1 - h_2$ pour les scripts *NasalityAutoMeasure* en version automatique, *NasalityAutoMeasure* en version manuelle, et notre propre *script* manuel pour chaque type de phonation (craqué, modal, soufflé)

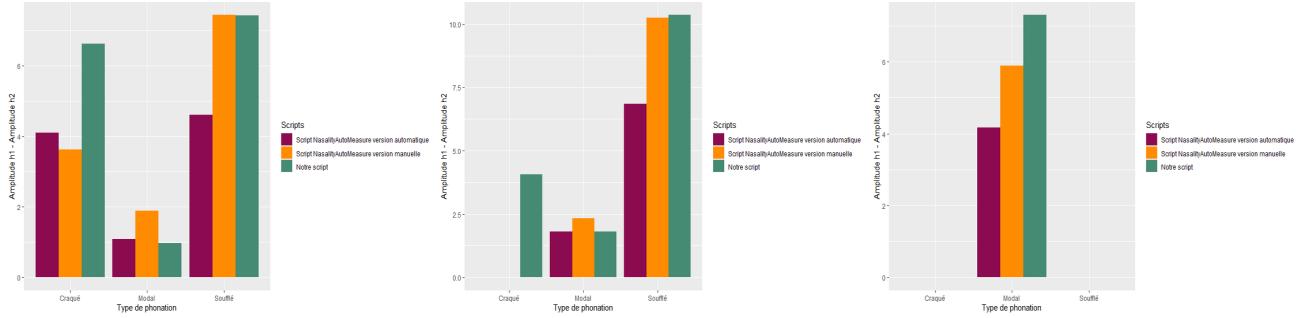


FIGURE 35 – Histogrammes groupés pour les locuteurs 10, 11 et 12 représentant la mesure $h_1 - h_2$ pour les scripts *NasalityAutoMeasure* en version automatique, *NasalityAutoMeasure* en version manuelle, et notre propre **script manuel** pour chaque type de phonation (craqué, modal, soufflé)

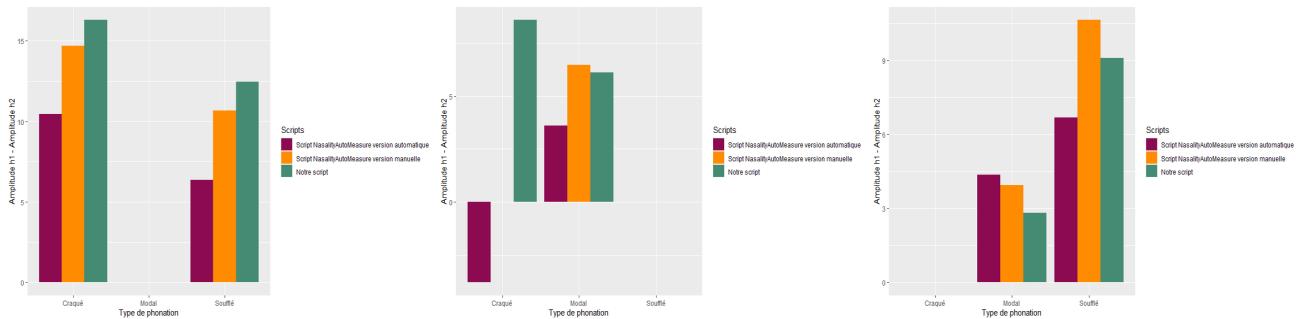


FIGURE 36 – Histogrammes groupés pour les locuteurs 13, 14 et 15 représentant la mesure $h_1 - h_2$ pour les scripts *NasalityAutoMeasure* en version automatique, *NasalityAutoMeasure* en version manuelle, et notre propre **script manuel** pour chaque type de phonation (craqué, modal, soufflé)

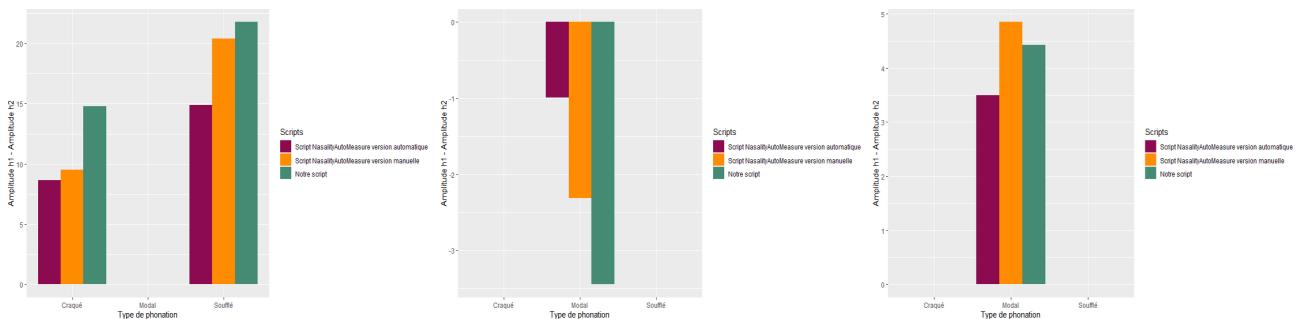


FIGURE 37 – Histogrammes groupés pour les locuteurs 16, 17 et 18 représentant la mesure $h_1 - h_2$ pour les scripts *NasalityAutoMeasure* en version automatique, *NasalityAutoMeasure* en version manuelle, et notre propre **script manuel** pour chaque type de phonation (craqué, modal, soufflé)

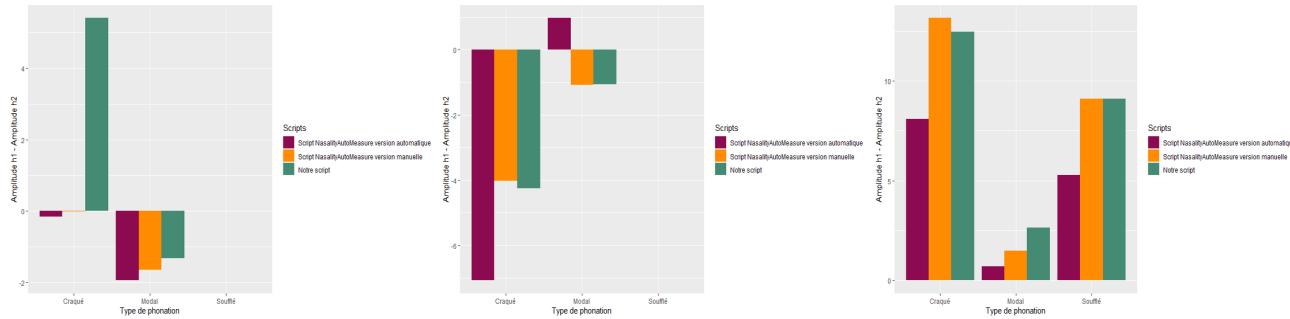


FIGURE 38 – Histogrammes groupés pour les locuteurs 19, 20 et 21 représentant la mesure $h_1 - h_2$ pour les scripts *NasalityAutoMeasure* en version automatique, *NasalityAutoMeasure* en version manuelle, et notre propre script manuel pour chaque type de phonation (craqué, modal, soufflé)

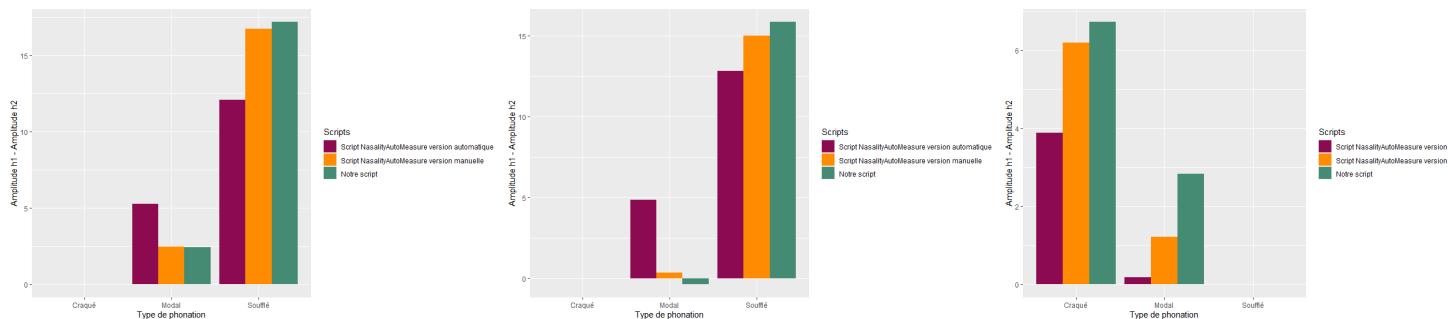


FIGURE 39 – Histogrammes groupés pour les locuteurs 22, 23 et 24 représentant la mesure $h_1 - h_2$ pour les scripts *NasalityAutoMeasure* en version automatique, *NasalityAutoMeasure* en version manuelle, et notre propre script manuel pour chaque type de phonation (craqué, modal, soufflé)

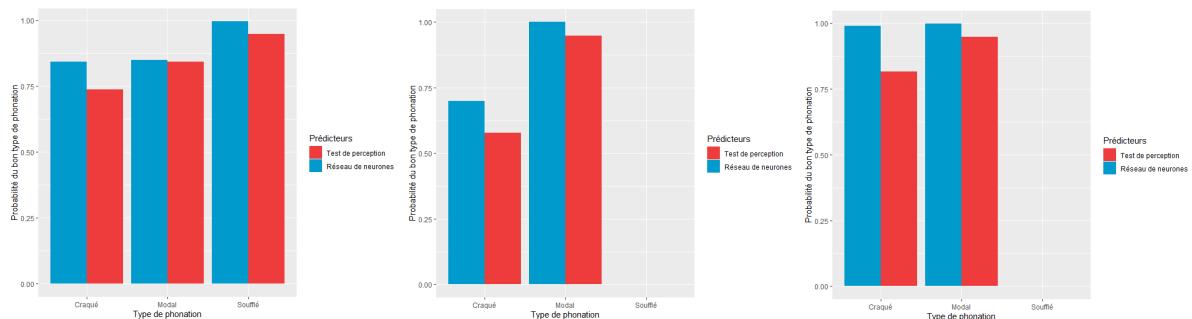


FIGURE 40 – Histogrammes groupés pour les locuteurs 1, 2 et 3 représentant le taux de probabilité pour le bon type de phonation du Réseau de neurones et le pourcentage de bonnes réponses pour le Test de perception pour chaque type de phonation (craqué, modal, soufflé)

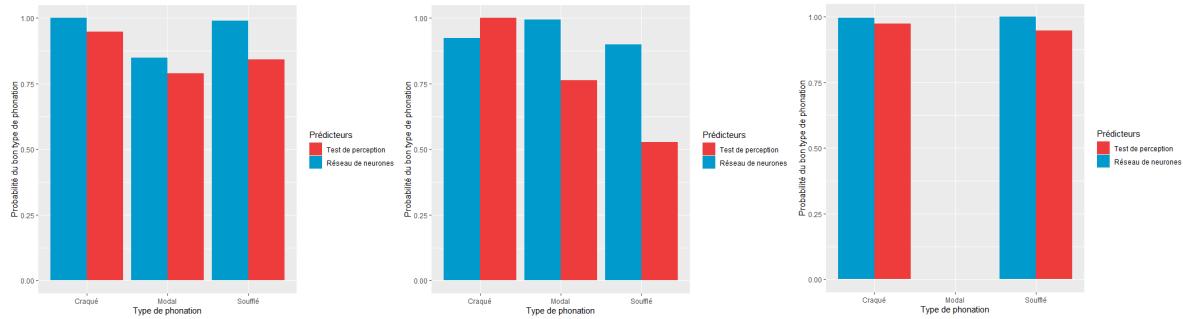


FIGURE 41 – Histogrammes groupés pour les locuteurs 4, 5 et 6 représentant le taux de probabilité pour le bon type de phonation du **Réseau de neurones** et le pourcentage de bonnes réponses pour le **Test de perception** pour chaque type de phonation (craqué, modal, soufflé)

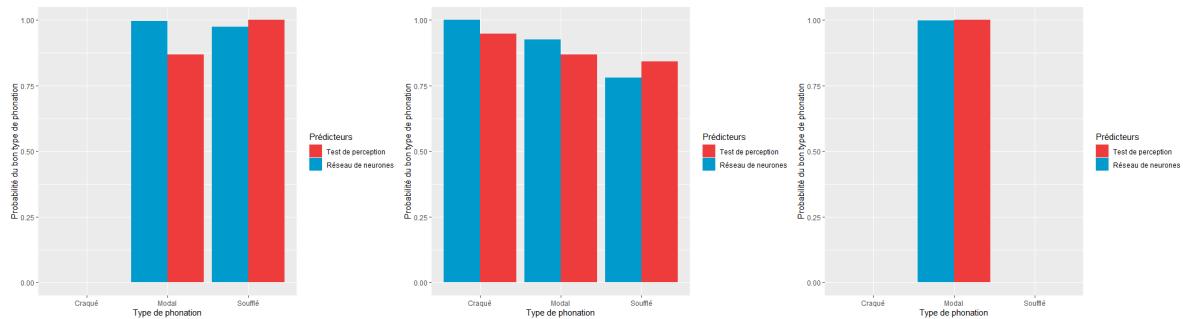


FIGURE 42 – Histogrammes groupés pour les locuteurs 7, 8 et 9 représentant le taux de probabilité pour le bon type de phonation du **Réseau de neurones** et le pourcentage de bonnes réponses pour le **Test de perception** pour chaque type de phonation (craqué, modal, soufflé)

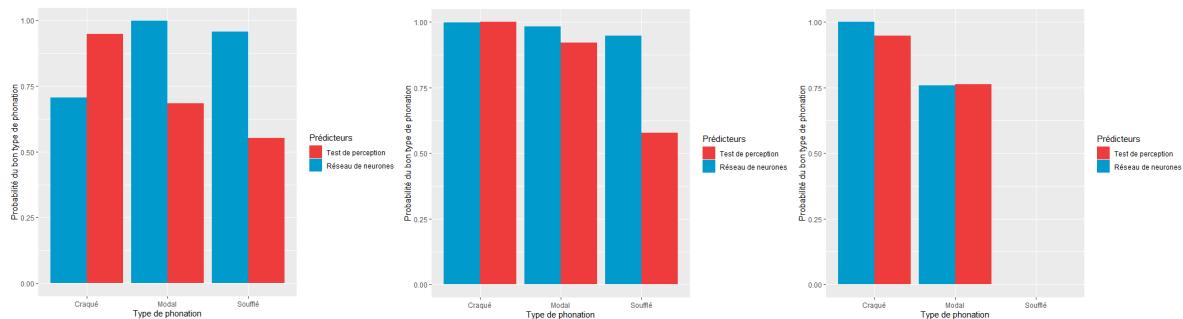


FIGURE 43 – Histogrammes groupés pour les locuteurs 10, 11 et 12 représentant le taux de probabilité pour le bon type de phonation du **Réseau de neurones** et le pourcentage de bonnes réponses pour le **Test de perception** pour chaque type de phonation (craqué, modal, soufflé)

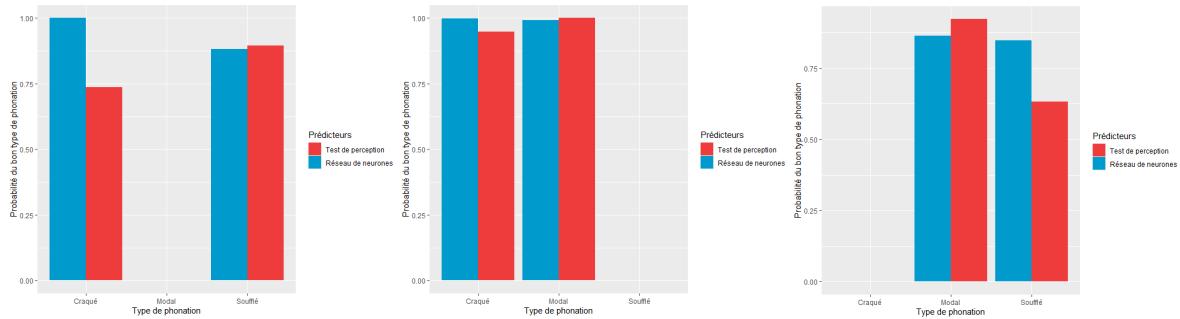


FIGURE 44 – Histogrammes groupés pour les locuteurs 13, 14 et 15 représentant le taux de probabilité pour le bon type de phonation du Réseau de neurones et le pourcentage de bonnes réponses pour le Test de perception pour chaque type de phonation (craqué, modal, soufflé)

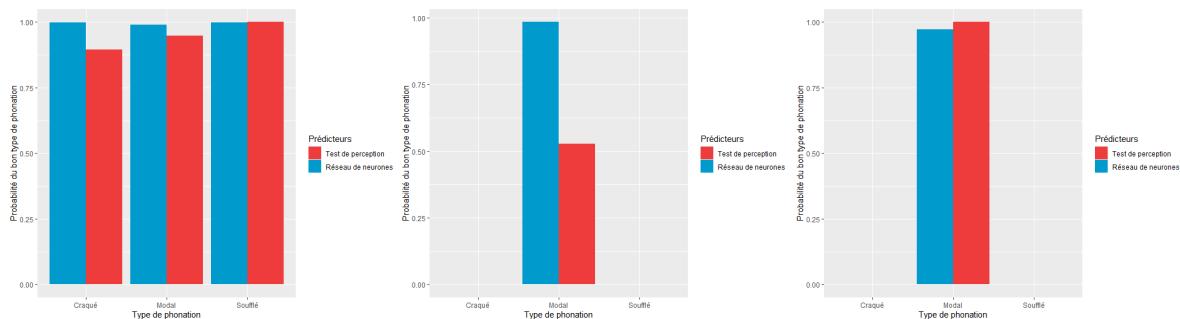


FIGURE 45 – Histogrammes groupés pour les locuteurs 16, 17 et 18 représentant le taux de probabilité pour le bon type de phonation du Réseau de neurones et le pourcentage de bonnes réponses pour le Test de perception pour chaque type de phonation (craqué, modal, soufflé)

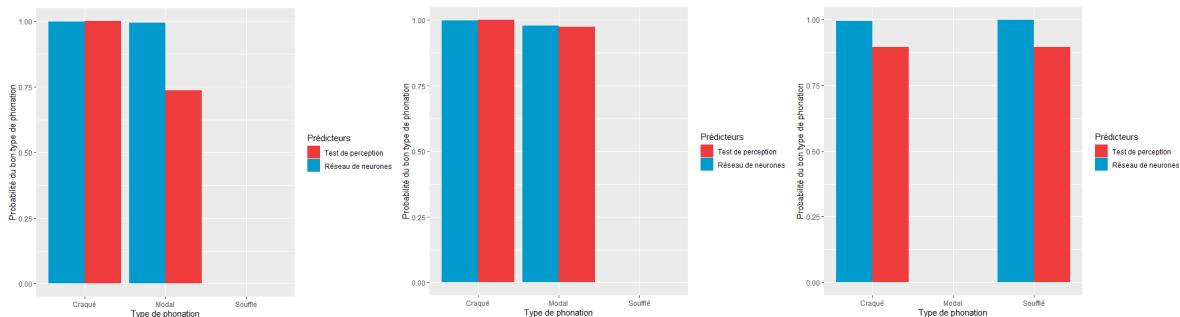


FIGURE 46 – Histogrammes groupés pour les locuteurs 19, 20 et 21 représentant le taux de probabilité pour le bon type de phonation du Réseau de neurones et le pourcentage de bonnes réponses pour le Test de perception pour chaque type de phonation (craqué, modal, soufflé)

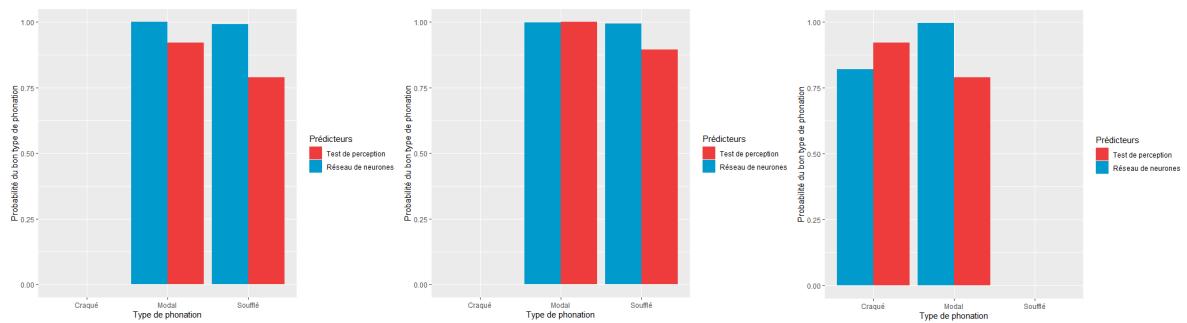


FIGURE 47 – Histogrammes groupés pour les locuteurs 22, 23 et 24 représentant le taux de probabilité pour le bon type de phonation du **Réseau de neurones** et le pourcentage de bonnes réponses pour le **Test de perception** pour chaque type de phonation (craqué, modal, soufflé)

Résumé du mémoire

Les types de phonation ont diverses implications phonétiques, pathologiques, culturelles... Et leur réalisation peut être dépendante de facteurs variant selon les individus.

Ce mémoire s'intéresse aux variations inter-locuteurs des différents types de phonations (voix modale, craquée et soufflée). Ces variations sont étudiées sous un angle acoustique avec la mesure $h_1 - h_2$, et sous un angle perceptuel avec l'évaluation des types de phonation par un groupe de participants et un réseau de neurones. L'objectif de cette étude est double : évaluer la présence de variation inter-locuteur pour les types de phonation, mais aussi tester la fiabilité des trois prédicteurs évoqués (humains, réseau de neurones et mesure acoustique).

Les résultats montrent une meilleure performance des humains et du réseau de neurones, tandis que la mesure acoustique $h_1 - h_2$ présente des résultats parfois erronés — partiellement dus à une mauvaise reconnaissance de la f_0 . L'autre facteur d'erreurs est l'analyse d'une seule partie de l'extrait, au lieu d'apprécier le type de phonation sur l'extrait entier — ce que font les humains et le réseau de neurones.