MÉMOIRE DE MASTER 2

Mention : Traitement Automatique des Langues

Spécialité : Recherche & Développement

Classification automatique de l'âge du locuteur. Influence des descripteurs, du sexe et des conditions d'enregistrement.

Par Audrey Gombault

Sous la direction de Mr Nicolas Audibert

Année universitaire 2018-2019

Université Paris Nanterre

N° étudiant : 37002194

SOMMAIRE

ENGAGEMENT DE NON-PLAGIAT

Je, soussignée Gombault Audrey, étudiante en Master 2 de Traitement Automatique des Langues (R&D) à l'Université de Paris-Nanterre, déclare être pleinement conscient(e) que le plagiat d'un document ou d'une partie de document publié sur toutes les formes existantes de support, y compris sur Internet, constitue une violation des droits d'auteur ainsi qu'une fraude caractérisée. En conséquence, je m'engage à citer explicitement, à chaque fois que j'en fais usage, toutes les sources que j'ai utilisées pour écrire ce mémoire.

Fait à Paris le 13/07/19

Audrey Gombault

Résumé

Cette étude porte sur la classification de l'âge du locuteur à partir d'extraits de parole, avec des descripteurs phonétiques seuls ou ajoutés à des paramètres cepstraux de type MFCC.

Nous évaluons les effets de variables telles que l'âge, le sexe et la condition d'enregistrement sur des paramètres phonétiques tels que la fréquence fondamentale, les valeurs de formants ou encore la durée des phones.

Pour cela, nous avons constitué un corpus constitué de deux sous-corpus enregistrés dans deux conditions différentes : une condition contrôlée, en présentiel avec le même matériel pour tous les locuteurs, et une condition moins contrôlée, pour laquelle les locuteurs s'enregistraient en ligne sur une plateforme d'enregistrement développée à cet effet. Les enregistrements recueillis auprès de 80 locuteurs âgés de 18 ans à 86 ans ont été répartis en sept classes d'âge, avec un maximum de quinze énoncés par locuteur.

La comparaison directe entre conditions d'enregistrement sur un sous ensemble de 13 locuteurs indique que globalement, la variabilité entre conditions pour un même locuteur est très inférieure à la variation entre locuteurs pour les mesures de durée, de F0 et de formants.

L'analyse d'un ensemble de descripteurs phonétiques en fonction de l'âge révèle une augmentation progressive des durées des phones et des syllabes, une augmentation de la fréquence fondamentale moyenne chez les locutrices de plus de 60 ans, et une baisse avec l'âge des formants F2 et F3 de la voyelle /a/ pour les locutrices de plus de 50 ans.

La comparaison des scores de classification obtenus par les algorithmes JRIP, J48 et SMO sur les données enregistrées en ligne indiquent que les performances les

meilleures sont obtenues avec l'algorithme SMO en séparant les données recueillies auprès des hommes et des femmes. Si les paramètres cepstraux permettent d'obtenir des taux de classification correcte nettement supérieurs à ceux des descripteurs phonétiques utilisés seuls (92% pour les femmes, 98% pour les hommes), une sélection de ces derniers conduit à une amélioration marginale des taux de classification allant jusqu'à 2,5

1 INTRODUCTION

Avec l'âge, des changements biologiques et physiologiques sont observés. En effet, il faut noter des changements au niveau du système respiratoire (capacité des poumons, affaiblissement des muscles respiratoires), au niveau du larynx (ossification des cartilages, atrophie des muscles laryngés), au niveau du squelette crânien qui continue de se développer à l'âge adulte ou encore au niveau de la langue dont les muscles vont s'atrophier (Schötz 2006). Dès l'adolescence, ces changements peuvent être observés, puisque le corps est en pleine croissance et il connait de fortes modifications avec la puberté.

Nous nous intéressons à la répercussion sur le plan acoustique et phonétique de ces changements. Nous essayons d'extraire ces paramètres phonétiques marqueur d'un vieillissement de voix, afin de pouvoir proposer une classification de l'âge.

Dans le domaine de la reconnaissance du locuteur, certains travaux se sont intéressés à la classification de l'âge auparavant, principalement avec des descripteurs qui ne permettent pas de savoir quelles caractéristiques de la parole ou physiques et physiologiques changent avec l'âge, les MFCC, coefficients cepstraux que nous vous présenterons dans ce rapport. Nous avons donc pour objectif de déterminer s'il est possible d'obtenir de meilleurs résultats de classification en utilisant des descripteurs phonétiques, et d'exploiter les résultats obtenus avec ces descripteurs phonétiques afin de mieux comprendre quels aspects de la production de la parole sont le plus affectés par les différences d'âge.

Dans un tout premier temps, nous allons vous présenter les avancées faites à la fois dans le domaine du vieillissement de la voix, par des études phonétiques, acoustiques

et perceptives, dans le domaine de la condition d'enregistrement, de la classification automatique et de la reconnaissance du locuteur.

Dans un deuxième temps, nous vous exposerons les méthodes employées, en termes de recueil des données, de préparation et de transformation des données, afin de pouvoir ensuite exploiter nos données.

Dans un troisième temps, nous vous présenterons les résultats des transformations effectuées qui nous ont permis de choisir le type de classification que nous allions appliquer, les descripteurs extraits de nos données que nous avons décidé de conserver pour la classification ainsi que les résultats de ladite classification.

Enfin, nous conclurons sur les résultats de la classification que nous aurons obtenus et nous vous présenterons de possibles prolongements de nos travaux.

2 ÉTAT DE L'ART

2.1 PHONÉTIQUE ET VIEILLISSEMENT DE LA VOIX

2.1.1 Signal acoustique

La parole est le résultat de mouvements de l'appareil phonatoire qui se découpe en deux parties, la source, les poumons et le canal, le conduit vocal, de cavités résonnantes (pharynx, cavité buccale et nasale) et des organes d'articulation (voile du palais, langues, lèvres…).

Nous pouvons distinguer différents sons produits par l'appareil phonatoire humain, d'un côté les voyelles, de l'autre les consonnes.

2.1.1.1 Voyelles

Les voyelles se caractérisent par un voisement et un passage non obstrué de l'air dans le conduit vocal.

Les voyelles sont des sons périodiques et présentent des « zones d'harmoniques renforcées appelées "formants" (Meunier 2007). Nous reviendrons sur la définition de ces formants un peu plus tard.

Les voyelles peuvent se caractériser de différentes façons, par leur ouverture (fermé à ouverte), leur antériorité (ou postériorité) et leur arrondissement (arrondi à non arrondi).

2.1.1.2 Consonnes

En français on distingue fricatives et occlusives. Les premières découlent d'un rétrécissement local du conduit vocal, tandis que les secondes résultent d'une obstruction totale du conduit en un point, puis de son ouverture « brusque » (Haton et al. 2006).

2.1.2 Spectre et spectrogramme

D'après (Cornut 2009), chaque son possède un spectre acoustique qui se découpe en deux parties ; la première étant le timbre vocalique, composée de la fréquence fondamentale et des deux premiers formants, qui est très similaire pour une même voyelle prononcée par différentes personnes, la deuxième étant le timbre extra-vocalique qui lui va donner des informations plus spécifiques et individuelles.

Il faut cependant noter qu'inclure la fréquence fondamentale dans le timbre vocalique est en contradiction la théorie source-filtre, présentée par (Vaissière 2015) et qui stipule que tous les sons voisés ont pour origine le « bourdonnement glottal » résultant du mouvement des plis vocaux. Ce bourdonnement évoluant dans un volume clos (conduit vocal), il aurait des résonnances naturelles.

Un spectrogramme est un ensemble de spectres et permet de représenter l'évolution du signal sonore, c'est une représentation des spectres successifs (fig. 2.1.2-1) calculés dans le temps.

Il réalise une représentation en trois dimensions : temps fréquence des composantes (harmoniques ou non) et intensité des composantes (Martin 2008).

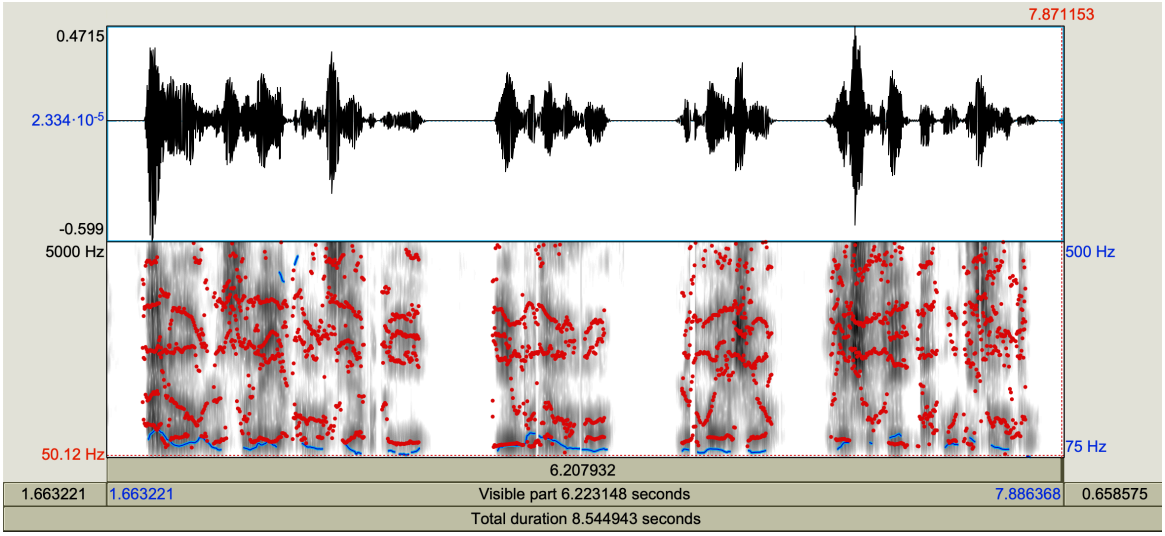


Figure 2.1.2-1: Exemple d'un spectrogramme affiché par PRAAT, les niveaux de gris représentent l'énergie associée aux différentes fréquences.

La majorité des paramètres qui peuvent être extraits du signal acoustique sont issus des représentations spectrales.

2.1.3 SPL (pression acoustique) et débit

(Schötz 2007) s'est intéressée à deux paramètres qui se sont avérés changer avec l'âge ; la durée du segment sonore et la pression sonore (notée SPL). La pression sonore est la pression mesurée entre l'onde sonore et la pression de l'air dans lequel le son est produit. Des sons forts produisent des ondes sonores avec une pression sonore forte et inversement. La SPL et la durée de segment ont tendance à augmenter avec l'âge chez les hommes et les femmes.

La durée des segments a tendance à se prolonger avec l'âge, et ce de manière plus marquée chez les hommes, ce qui aura également une influence sur le débit.

Le débit correspond au « calcul du nombre de mots par minute, du nombre de syllabes par minute ou par seconde, évalué sur tout ou partie du temps de parole, pauses

comprises ou non. Lorsque le débit est calculé à partir du temps de parole dont est exclue la durée des pauses, il devient un indice de la vitesse d'articulation » (Colletta, Pellenq, and Rousset 2008). On parle de débit articulatoire (Simon 2007).

Le débit a tendance à diminuer avec l'âge, de manière plus distincte chez les hommes que chez les femmes, pour qui il n'y a parfois pas de changement (Schötz 2006).

2.1.4 Fréquence fondamentale

La fréquence fondamentale correspond à la vitesse des cycles d'ouverture/fermeture de la glotte pendant la production des sons voisés. Sur Praat, elle est extraite grâce à la méthode d'autocorrélation (Boersma 1993).

Pour la fréquence fondamentale, (Torre III and A. Barlow 2009), en se basant sur de précédentes études (Barbaranne J. Benjamin, 1981, 1982 et 1986 ou Linville et Fisher 1985), ont mis en évidence le consensus concernant la différence entre la hauteur de la voix des femmes (moyenne 190Hz voire 220Hz) et celle de la voix des hommes (accord à 120-130Hz). Ils ont de plus observé différentes observations quant à la variation de la fréquence fondamentale avec l'âge : « *There also is some disagreement in the literature regarding age-related changes of F 0 for the two sexes. For instance, some have reported a slight increase in F 0 for men, and a significant decrease for women (Mueller, 1985, 1997; Mueller, Sweeney, Baribeau, 1984; Nishio, Niimi, 2008; Russell, Penny, Pemberton, 1995), while others have noted a significant increase for men, and only a slight decrease for women (Benjamin, 1981, 1982; Hollien, Shipp, 1972; Honjo, Isshiki, 1980; Mysak, 1959; Ramig et al., 2001).* ». (Schötz 2007)] a trouvé des changements significatifs chez les deux sexes, ainsi, la F0 chez les femmes descendrait jusqu'à 50 ans puis se stabiliserait. Pour les hommes, on observerait une baisse jusqu'à 40-50 ans puis une forte hausse.

2.1.5 Formants

Le son émis par les plis vocaux a donc une fréquence fondamentale (présentée en section) et des harmoniques qui sont des multiples de la fréquence fondamentale. Les formants correspondent à des zones d'harmoniques renforcées et donnent des informations sur la forme des cavités qui les ont créées (Vaissière 2015).

Les derniers paramètres étudiés par (Schötz 2007) est la fréquence de résonnance du conduit vocal, les formants, pour les voyelles, dont les trois premiers sont F1, F2 et F3, qui correspondent généralement à l'ouverture, l'antériorité et l'arrondissement de la voyelle.

(Schötz 2007) reporte que de l'étude de Endres et al. 1971, les fréquences de formants ont tendance à diminuer avec l'âge, pour les femmes comme pour les hommes, et cela serait dû à l'allongement du conduit vocal. C'est ce qu'elle a pu démontrer lors de son étude pour certaines voyelles seulement : la valeur de F1 baisse pour /ɛ:/ (et dans /y:/ pour les femmes) mais reste stable pour les autres voyelles. La valeur de F2 reste stable pour /y:/ mais augmente pour /ɑ:/ et /ɛ:/ mais diminue pour /a/ et /u:/ avec augmentation puis pic à 40 ans.

Rastatter et al., 1990 ; Rastatter et al., 1997 auraient observé une tendance à la centralisation des voyelles mais ce ne serait pas systématique pour tous les locuteurs.

Linville et al., 2001 ont observé une baisse significative des formants F1, F2 et F3 chez les femmes et une baisse de du formant F1 seulement chez les hommes (Schötz 2007).

2.1.6 Centre de gravité spectral

Le centre de gravité spectral (CGS) est donc une moyenne pondérée des amplitudes, une moyenne de l'énergie dans le spectre d'un son. Plus il est élevé plus il y a de l'énergie dans les hautes fréquences. Le CGS est un corrélat acoustique du lieu

articulation des consonnes, plus il est haut, plus le point d'articulation a tendance à être avancé dans la cavité buccale.

« The spectral centroid is commonly associated with the measure of the brightness of a sound. This measure is obtained by evaluating the "center of gravity" using the Fourier transform's frequency and magnitude information. The individual centroid of a spectral frame is defined as the average frequency weighted by amplitudes, divided by the sum of the amplitudes »(Nam 2001).

D'après l'étude de (Schötz 2007), certains sons seraient plus porteurs d'informations sur l'âge. Il s'agirait ainsi des trois consonnes occlusives sourdes /p/ /t/ /k/ qui correspondent aux consonnes qui n'ont pas d'énergie dans les moyennes et hautes fréquences et de la fricative sourde /s/.

2.1.7 ZCR

Le Zero-Crossing-Rate (ZCR) est la mesure du nombre de fois où l'amplitude des signaux de parole passe par zéro, dans une fenêtre temporelle donnée. C'est une simple mesure de la fréquence d'un signal (Nam 2001).

Il permet de distinguer les segments voisés et non-voisés de la parole (Santini 2016), car de hautes fréquences impliquent un ZCR haut, et de basses fréquences impliquent un ZCR bas. Et il existe une corrélation entre le ZCR et l'énergie, par conséquent, plus le ZCR est bas, plus le son est voisé et plus le ZCR est haut moins le son est voisé (Bachu et al. 1978).

Il faut également noter que le bruit de fond a une incidence sur le ZCR, surtout si le SNR (rapport signal sur bruit) est faible car le bruit de fond sera plus difficilement identifiable (Bachu et al. 1978).

Ces différents paramètres mis en avant dans la littérature nous seront utiles dans l'application de méthodes utilisées en traitement automatique des langues, telles que de la reconnaissance du locuteur et de la classification.

2.2 LA RECOLTE DE DONNEES EN LIGNE

Common Voice fait partie de l'initiative Mozilla est un grand corpus constitué de données de parole recueillies en ligne sur une plateforme prévue à cet effet. Il a été conçu dans l'objectif de fournir un jeu de données assez grand pour que les technologies de la reconnaissance vocale puissent s'améliorer.

2.3 RECONNAISSANCE DU LOCUTEUR ET CLASSIFICATION

La tâche de base de la reconnaissance du locuteur est de déterminer si un segment sonore a été prononcé ou non par un locuteur donné.

La classification est, dans notre cas, peut ressembler à une étape du Data Mining ou fouille de données qui s'inscrit dans un grand ensemble de techniques de traitement et stockage de l'information, appelé Extraction des Connaissances à partir des Données (ECD). Cet ensemble prend en compte l'étape de recueil des données jusqu'à l'étape d'interprétation et évaluation des connaissances extraites.

Une fois les données recueillies, une première étape est la sélection des données qui seront pertinentes et donnerons les données transformées. Ces dernières seront ensuite préparées puis transformées afin de pouvoir passer à l'étape du data-mining ou fouille de texte dont seront extrait des patrons, classifications qui seront interprétées et évaluées pour devenir des connaissances.

La fouille de texte est « un domaine multidisciplinaire, mélangeant la technologie des bases de données, l'intelligence artificielle, l'apprentissage automatique, les réseaux de neurones, les statistiques, la reconnaissance de formes, les systèmes à bases de

connaissances, l'acquisition de connaissance, les systèmes de recherches d'informations, l'informatique haute-performance, et la visualisation de données » (Jollois 2003). Nous nous intéresserons donc plus particulièrement au domaine de la classification automatique.

La classification est une tâche qui consiste à « associer une "classe" à chaque donnée d'entrée » (Tellier n.d.).

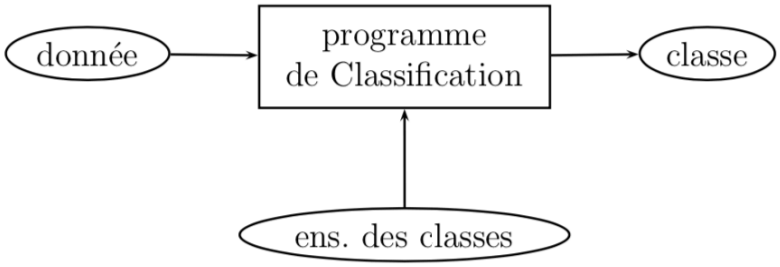


Figure 2.1.7-1 : Schéma général de la tâche de classification (Tellier n.d.).

Ces classes doivent répondre au critère de « compacité » ou de « ressemblance », c'est-à-dire que les éléments de chaque classe doivent être le plus similaire possible, et les classes doivent également répondre au critère de « séparabilité », c'est-à-dire que la distance qui les sépares doit être assez conséquentes pour qu'elles ne se confondent pas (Tellier n.d.).

Les données d'entrées peuvent être de différents types, celles que nous utiliserons seront des valeurs numériques qui correspondent à des paramètres acoustiques extrait du signal acoustique.

En classification, se distinguent deux méthodes d'apprentissage ; la méthode supervisée, pour laquelle est donnée au programme les sorties possibles, la classification est guidée et la méthode non supervisée, pour laquelle le programme doit regrouper en choisissant lui-même les paramètres qui ont une importance et qui déterminent la distance entre deux éléments. La classification non supervisée par donc de l'hypothèse que les

éléments qui constituent les données présentent effectivement des différences notable et significatives, qui permettent de séparer ces éléments en plusieurs groupes alors que pour la classification supervisée, l'algorithme d'apprentissage va extraire les règles de groupes déjà formé.

2.3.1 Classification supervisée

2.3.1.1 Classe majoritaire

L'algorithme de classe majoritaire est à envisager comme une « baseline » ou record à battre (Tellier n.d.) et il constitue un premier exemple de classification par apprentissage supervisé. Cet algorithme ne revoie qu'une seule valeur, il est alors préférable de choisir la classe la plus représentée dans le jeu de données. Pour exemple, dans un contexte d'apprentissage automatique, on peut vouloir identifier ce qui est un adjectif et ce qui n'en est pas un (on a alors deux classes ADJ|NON-ADJ), l'algorithme de classe majoritaire dans ce cas ne renverrait que des « NON-ADJ ».

2.3.1.2 K plus proches voisins

(Rabiner et al. 1979) ont utilisé la méthode supervisée des KNN (k-nearest neighbors) sur des données basées sur une analyse en LPC (linear predictive coding), très populaire en traitement du signal et de la parole et qui sert à représenter l'enveloppe spectrale d'un signal. Cette technique consiste à prédire le signal à un instant n à partir des p échantillons précédents.

(Rabiner et al. 1979) ont fait varier les paramètres de leur KNN, en faisant varier K (de 1 à 4), C, le nombre de candidats ordonnés considérés (de 1 à 5) et l, le nombre de modèles par mot (2 à 12). Les meilleurs résultats s'obtiennent lorsque K diminue, que le nombre de modèles augmente, que l'ensemble de mot est prédéfini plutôt qu'aléatoire et

que le nombre de candidats augmente, lorsque K=1, I=12 et C=5, le taux de reconnaissance est de 97,9% contre 58,3% quand K=5, I=2 et C=1.

2.3.2 Classification non supervisée

2.3.2.1 Support Vector Machine (SVM)

Les SVM sont très utilisés dans la reconnaissance de l'âge et du sexe du locuteur. Ils sont très sensibles aux données bruyantes car ils donnent le même poids à tous les points de données (Phuoc et al. 2010). Pour pallier ce problème, il est possible d'attribuer une valeur d'appartenance factice (« a fuzzy membership value ») comme poids à chaque point de données. (Phuoc et al. 2010) ont réalisé deux expériences, la première attribuant une valeur d'appartenance égale à tous les vecteurs de paramètres d'une même classe et la seconde des valeurs d'appartenances différentes pour chaque vecteur de paramètres d'une même classe. Les deux expériences ont obtenu et de meilleurs résultats que la baseline qui était de 47.11% (respectivement 48.61% et 48.8%) pour la détection du sexe comme de l'âge, mais le taux de détection de l'âge reste très faible (<50%) et non satisfaisant. Leur classification a porté sur six classes, quatre classes d'âge et deux classes de sexe.

2.3.2.2 GMM (gaussian mixture model)

Le GMM est un modèle probabiliste qui permet de représenter des sous-populations d'individus dont la distribution suit une loi normale au sein d'une plus grande population. Le modèle mixte va apprendre cette sous-population, il s'agit en quelques sortes d'un apprentissage non-supervisé.

« A GMM is used in speaker recognition applications as a generic probabilistic model for multivariate densities capable of representing arbitrary densities, which makes it well suited for unconstrained text-independent applications. » (C. Platt 1999).

2.3.3 Paramétrisation du signal de parole :

On peut choisir des paramètres de nature phonétique (F0, formants, durée) présentés en section 2.1, mais la plupart des études de classification ont eu recours à des paramétrisation du signal qui peuvent être extraites de façon intégralement automatique.

2.3.3.1 MFCC (Mel Frequency Cepstral Coefficient)

Les MFCC représentent les coefficients cepstraux. Les bandes de fréquence de ce spectre sont espacées selon l'échelle de Mel qui est une échelle logarithmique qui s'applique au découpage de fréquence.

Les coefficients sont représentés dans une succession de fenêtres temporelle equiréparties avec une période d'échantillonnage constante.

Ces coefficients sont obtenus grâce à la transformation du spectre en cepstre, grâce à une transformée de Fourier inverse du logarithme de la transformée de Fourier et ce découpage du cepstre en fonction de l'échelle Mel.

(Davis et Mermelstein 1980) rappellent que dans une étude antérieure, les MFCC avaient été de bonnes représentations des informations consonantiques. Une représentation comprenant seulement 2 coefficients cepstraux avait déjà permis de reconnaitre correctement à 96% des mots phonétiquement similaires ("stick," "sick," "skit," "spit," "sit," "slit," "strip," "scrip," "skip," "skid," "spick," et "slid ").

(Moritz et al. 2016) avaient eu recours à des paramètres MFCC comprenant 13 coefficients cepstraux pour leur GMM-HMM.

(Schötz 2006) réutilise l'étude de Minematsu et al. (2002 a, b) qui proposaient une classification de l'âge perçu grâce aux MFCC comme paramètres acoustiques pour une GMM et une distribution normale (méthodes de classification : LDA et ANN). Les voix les plus vieilles étaient correctement classifiées dans 90.9% avec la méthode LDA.

(Deshpande, Singh, Nam 2001) ont utilisé douze coefficients cepstraux pour appliquer les classifications SVM, KNN et un GMM. Le meilleur résultat qu'ils aient obtenu en la classification à trois classes est pour le KNN (75%). Pour la classification à deux classes de la musique classique, c'est le SVM qui obtient les meilleurs résultats (90%).

Pour leur classification pour l'identification de l'âge et du sexe, (Shepstone, Tan, Holdt Jensen 2013) ont utilisé treize coefficients cepstraux et ont également utilisé les dérivées première et seconde pour obtenir un total de trente-neuf coefficients cepstraux. La précision du sous-système acoustique n'est cependant que de 49.9 %.

3 MATÉRIAUX ET MÉTHODES

3.1 RECUEIL DES DONNÉES

Afin de mener à bien notre étude, nous avons dû constituer un corpus qui répondrait à nos exigences, tant en contenu qu'en métadonnées. En effet, nous avons trouvé très peu de données exploitables dans le cadre de nos recherches car il manquait souvent des informations sur l'âge des locuteurs. Nous avons donc constitué un corpus que l'on pourrait diviser en deux sous-corpus ; le premier est constitué de données récupérées en ligne, le second est composé de données récupérées directement auprès de locuteurs à l'aide d'un microphone et d'une carte son.

Quel que soit le support de recueil, les locuteurs devaient prononcer quinze énoncés, choisis en fonction des précédentes lectures sur les phones les plus porteurs d'informations sur l'âge du locuteur. Certains de ces énoncés se retrouvent dans des corpus que nous avons également à disposition mais qui ne sont pas initialement prévus pour la reconnaissance de l'âge du locuteur, tels que les corpus MonPaGe (Lévêque et al. 2016), conçu pour l'évaluation de la parole francophone de patients présentant des signes de troubles moteurs de la parole, PTS Vox qui a été constitué dans le cadre du projet

VoxCrim[1], un projet de reconnaissance du locuteur dans le domaine criminalistique, ou encore des énoncés ont pu être extrait de la fable d'Esope, *La bise et le soleil*.

Les énoncés choisis pour constituer notre corpus sont les suivants :

01 Mélanie vend du lilas.

02 L'oiseau s'envole au premier bruit qui l'effraie.

03 L'homme sur le ponton porte un blouson blanc.

04 Anne-Marie et moi allons à la mer.

05 Papy-Louis et Papa vivent dans le sud de l'île de Tipapa.

06 Papy-Louis loua six hectares de terre au sud de l'île.

07 Papa loua un six pièces sur le site web de Tipapa.

08 Pour les transports en ville, Papa loua deux vélos et pour la ferme, Papy-Louis loua deux tracteurs.

09 À Tipapa, le système juridique est spécial : il y a six lois principales.

10 Le peuple de Tipapa propose des lois pendant les six premiers jours de l'année.

11 A Tipapa, le sénat rédige la loi des finances et le congrès du sud s'occupe des lois civiles.

12 Papy-Louis et Papa sont heureux dans le sud de l'île.

13 Ils doivent faire attention à ne pas faire tomber leurs oeufs dans la mer.

14 Je m'approchais du bord de la fenêtre pour regarder dans la rue.

15 Alors le soleil a commencé à briller et au bout d'un moment, le voyageur, réchauffé a ôté son manteau.

Nous allons alors vous présenter ces données sur lesquelles nous avons travaillé.

[1] https://anr.fr/Projet-ANR-17-CE39-0016

3.1.1 Récolte des données en ligne

Dans l'objectif de récolter un jeu de données important et de pouvoir effectuer des analyses phonétiques, nous avons développé un site codé en PHP, HTML et JAVASCRIPT, à partir d'un exemple d'utilisation de Recorder.js (https://github.com/mattdiamond/Recorderjs) qui s'appuie sur la Web Audio API (https://developer.mozilla.org/fr/docs/Web/API/Web_Audio_API), sur lequel les utilisateurs peuvent s'enregistrer en train de lire les énoncés sélectionnés. Les énoncés sont présentés aux utilisateurs dans un ordre aléatoire qui change d'un utilisateur à un autre. Les enregistrements recueillis sont stockés sur le serveur au format Wav.

Les utilisateurs avaient le choix de s'enregistrer sur ordinateur, tablette ou smartphone, par le biais de microphone intégré, de kit main-libre voire de microphone externe. Le site web a été diffusé par le biais des réseaux sociaux et par le biais des connaissances afin qu'il puisse être partagé à plus grande échelle.

Lorsque l'utilisateur se connecte sur le site, il remplit un formulaire qui nous permet de connaitre : son âge, son sexe, sa langue maternelle, s'il fume, s'il présente des troubles de lecture / parole, le matériel utilisé pour s'enregistrer. Ce sont des informations qui nous intéressaient du fait qu'elles peuvent avoir une influence sur la qualité et le contenu des enregistrements recueillis et sur nos analyses suivantes, et qui, une fois renseignées, sont écrites dans un fichier texte rattaché à l'utilisateur. Ce fichier texte nous indique également dans quel ordre ont été soumis les énoncés à l'utilisateur.

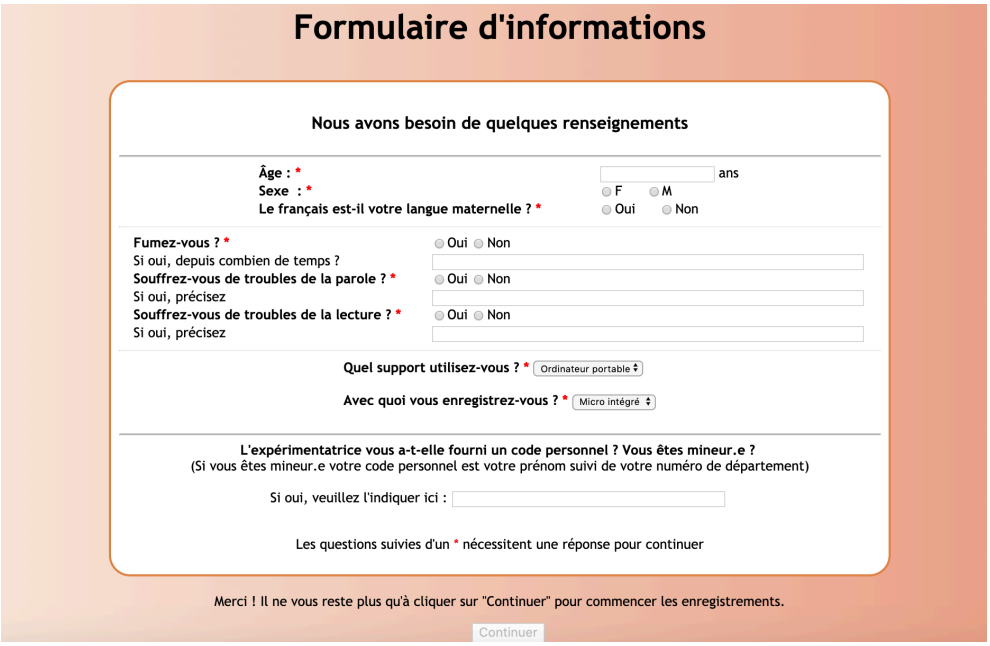


Figure 3.1.1-1 : Formulaire d'informations soumis à l'utilisateur.

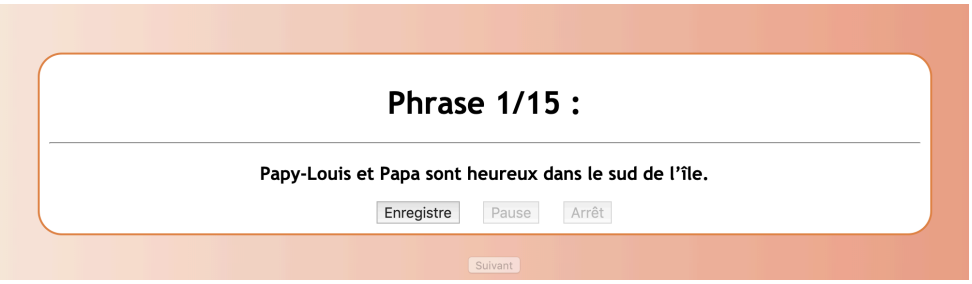


Figure 3.1.1-2 : Page d'enregistrement soumise à l'utilisateur.

3.1.2 Récolte des données en présentiel avec microphone et carte-son

Afin d'avoir une condition de référence, nous avons constitué un corpus de données enregistrées à l'aide d'un microphone (AKG 420) et d'une carte-son (Roland UA-22) empruntés à l'Ilpga.

Ces données ont généralement été enregistrées au domicile de la personne, dans un environnement silencieux. Certains locuteurs ont été enregistrée dans une grande salle, malgré le silence la voix des locuteurs résonne et cela détériore la qualité de l'enregistrement, les données de six locuteurs n'ont pas pu être exploitées de ce fait.

Nous avons également perdu les données d'une dizaine de locuteurs du fait d'un ventilation bruyante qui vient parasiter l'enregistrement.

Certains sujets enregistrés avec le microphone et la carte-son ont été invités à s'enregistrer ensuite ou au préalable sur le site d'enregistrement en ligne, de manière à ce que l'on puisse comparer la situation contrôlée (en présentiel) et la situation plus libre (recueil en ligne). Ces locuteurs ont été "marqués" par un code que nous leur avons attribué et qui se retrouve dans leurs fichiers "results" associés à leurs enregistrements, afin que nous puissions faire le lien entre les deux, le nom de locuteur étant attribué à partir de l'adresse IP pour les enregistrements en ligne.

Cet échantillon de locuteurs est composé de neuf femmes (18 à 58 ans) et quatre hommes (29 à 64 ans), identifiés sous les codes « MARINE », « MOI », « PAU », « TAT », « CLA », « MAN », « MAM », « VERO », « SYL », « ULR », « HER », « PAP » et « XAV ».

Nous avons effectué une démarche auprès de plusieurs lycées et collège afin de mettre en place des autorisations parentales pour que l'on puisse exploiter enregistrer certains élèves ou diffuser le site au sein de l'établissement, mais nous n'avons malheureusement pas obtenu de réponse positive.

3.2 PRÉPARATION DES DONNÉES RÉCOLTÉES

Les données qui ont demandé le plus de préparation étaient les données enregistrées au moyen du microphone et de la carte son car il a fallu découper manuellement les différents énoncés d'un enregistrement global.

Afin de préparer nos données à des traitements acoustiques et phonétiques, nous avons eu recours aux outils en ligne BAS ("Bavarian Archive for Speech Signals"), développés par l'infrastructure CLARIN-D. Il s'agit d'un ensemble de services tels que la segmentation ou l'étiquetage automatique de signaux de parole, de conversion de

graphèmes vers phonèmes et bien plus[2]. Nous nous sommes intéressés au service WebMAUS Basic, une version en ligne de l'outil MAUS ("Munich AUtomatic Segmentation") (Schiel 1999). Il s'agit d'un outil d'alignement forcé qui prend un fichier audio (.wav) en entrée, ainsi que sa transcription orthographique associée (.txt) en entrée et fournit des TextGrids lisibles par PRAAT en sortie que nous utiliserons pour faire des analyses acoustiques.

Nous avons donc apparié automatiquement les fichiers .wav avec leur transcription orthographique dans un fichier .txt à l'aide d'un script Python, afin que WebMAUS Basic nous fournisse les TextGrids correspondants et que nous puissions lancer les analyses acoustiques.

Cependant, MAUS ayant été développé par des chercheurs allemands, la phonétisation n'est pas optimisée pour le français et nous nous attendions à quelques erreurs. Ainsi en contrôlant de plus près les TextGrids générés, nous avons pu observer que certaines phonétisations étaient approximatives. Par exemple l'énoncé « L'oiseau s'envole au premier bruit qui l'effraie. » a été transcrit en alphabet SAMPA [lwaz a~vOl o pR@mje bRHi ki lfRE] au lieu de [lwazo sa~vOl o pR@mje bRHi ki lefRE] ou encore l'énoncé "Ils doivent faire attention à ne pas faire tomber leurs œufs dans la mer." où le segment "leurs œufs" est transcrit [l9Rof]. Cette dernière transcription a de plus soulevé un nouveau problème. Nous avons imaginé une solution afin de résoudre ce problème, il aurait s'agit de mettre en place un apprentissage automatique en repérant quelques énoncés de chaque prononciation et de comparer le centre de gravité spectrale par exemple pour le /f/

[2] https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface

car il s'agit d'une occlusive sourde et ce sera donc plus facilement repérable dans le signal acoustique.

Afin de résoudre le souci d'optimisation de la transcription, nous avons décidé de passer par le service G2P (« Grapheme to phoneme ») disponible sur le site BAS et provenant de l'outil « Balloon » développé initialement pour l'allemand et l'anglais et qui fait donc de la conversion graphèmes vers phonèmes (D. Reichel 2012). Il s'agit donc de l'étape intermédiaire entre la transcription orthographique et la TextGrid proposée par WebMAUS Basic, puisque le fichier de sortie est un fichier .par qui va contenir la transcription SAMPA de l'énoncé, afin de pouvoir utiliser WebMAUS General qui prend, en entrée, le fichier audio .wav et la transcription en phonèmes, le fichier .par.

Nous avons donc lancé la conversion graphèmes sur nos quinze énoncés et avons pu observer quelques erreurs de transcription. Nous avons donc corrigé ces erreurs (Tab. 3.1.2-1 et 3.1.2-2.), qui sont en réalité des phonèmes oubliés ou ajoutés (liaisons incongrues).

Tableau 3.2-1: Version non-corrigée d'une sortie .par

KAN: 0 lwaz
KAN: 1 a~vOl
KAN: 2 o
KAN: 3 pR@mje
KAN: 4 bRHi
KAN: 5 ki
KAN: 6 lfRE

Tableau 3.2-2 : Version corrigée d'une sortie .par (les caractères en gras ont été ajoutés car manquants dans la version de base)

KAN: 0 lwaz**o**
KAN: 1 **s**a~vOl
KAN: 2 o
KAN: 3 pR@mje
KAN: 4 bRHi
KAN: 5 ki
KAN: 6 l**e**fRE

Pour obtenir un fichier .par par enregistrement, nous avons de nouveau réalisé un script PYTHON (prep_file_for_WEBMAUS_general_enligne.py, Annexe 1)qui va associer la version corrigée de la transcription en phonèmes à son fichier audio.

WebMAUS General renvoie une TextGrid comme l'illustre la figure 3.1.2-1., avec la première ligne d'annotation qui est la transcription orthographique, la deuxième qui représente l'annotation KAN-MAU, qui est une segmentation en mots mais ces derniers sont transcrits en SAMPA, et la transcription SAMPA dans la dernière ligne d'annotation.

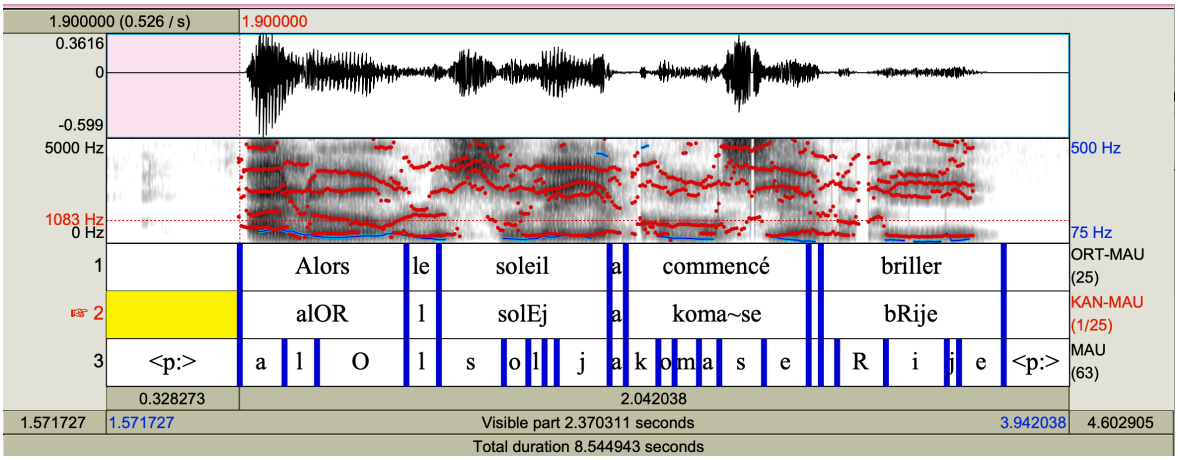


Figure 3.1.2-1: Exemple de visualisation d'un fichier son avec sa TextGrid associée, comprenant la ligne de transcription orthographique (ORT-MAU), la phonétisation en SAMPA au niveau des mots (KAN-MAU) et la segmentation en phones (MAU), également en SAMPA.

3.3 TRANSFORMATION

Une fois que nous avions préparé nos données, nous avons pu passer à l'étape des analyses acoustiques et phonétiques. Pour ce faire, nous avons utilisé l'outil PRAAT, qui permet d'analyser, synthétiser et manipuler les sons de parole. PRAAT permet l'analyse spectrale, de formants, de pitch, d'intensité mais aussi d'annoter des segments sonores par la génération de fichiers d'annotation au format TextGrid.

Nous vous présentons, dans cette partie, les traitements effectués par PRAAT sur nos données.

3.3.1 Traitement acoustique

Un des avantages de PRAAT est qu'il possède son propre langage de programmation et permet donc d'appliquer des scripts d'analyse sur de grands jeux de données et donc d'automatiser ces analyses.

Nous avons donc appliqué le script PRAAT d'extraction de paramètres acoustiques à partir de phones présélectionnés, développé par Nicolas Audibert, qui requiert en entrée le répertoire contenant les TextGrids, le répertoire contenant les fichiers audio, une liste contenant les noms des fichiers wav associés aux noms des fichiers TextGrid que l'on souhaite analyser (create_pairedfilelist.py Annexe 2), le nom du fichier de sortie, la tier (ligne d'annotation) qui contient les phones (en SAMPA), le fichier de paramètres que l'on souhaite connaitre, le fichier contenant les positions relatives des points de mesures cibles pour l'extraction et enfin la liste des phones qui nous intéressent.

Les paramètres que nous souhaitions extraire sont les valeurs de fréquences fondamentales et de formants (pour les voyelles), les valeurs de centres de gravité spectraux (CGS, principalement pour les consonnes occlusives et fricatives), le taux de passage par zéro (Zero-Crossing-Rate, ZCR), le rapport harmonique sur bruit (Harmonics-to-Noise Ratio, HNR) et l'intensité.

3.3.2 Modification du fichier de sortie (python et r)

Le fichier de sortie est au format .txt et contient un tableau des valeurs extraites sur lesquelles nous appliquerons ensuite des analyses statistiques.

textgrid_file	label	start_time	end_time	duration(s)	previousLabel	followingLabel	mean_F0(Hz)	F0_pt1_10%(Hz)
32_F_CEC_phrase05.TextGrid	a	0.64	0.71	0.06999999999999995	p	p	234.49837045982227	229.20516330974735
32_F_CEC_phrase05.TextGrid	i	0.82	0.89	0.07000000000000006	p	l	255.34917735464663	282.4035609616222
32_F_CEC_phrase05.TextGrid	u	0.97	1.04	0.07000000000000006	l	i	229.0731145664574	229.35427224366575
32_F_CEC_phrase05.TextGrid	i	1.04	1.21	0.16999999999999993	u	e	288.77099050374244	259.32698327913965
32_F_CEC_phrase05.TextGrid	e	1.21	1.33	0.1200000000000001	i	p	231.91584650887808	454.64475396602944
32_F_CEC_phrase05.TextGrid	a	1.43	1.5	0.07000000000000006	p	p	215.65526536209086	209.55808187028285
32_F_CEC_phrase05.TextGrid	a	1.63	1.91	0.28	p	v	251.65710455275112	225.76224179578298
32_F_CEC_phrase05.TextGrid	i	2	2.15	0.1499999999999999	v	v	237.28965018359992	205.06424426567503
32_F_CEC_phrase05.TextGrid	a~	2.34	2.44	0.10000000000000009	d	l	199.25635684693503	208.42115758624843
32_F_CEC_phrase05.TextGrid	@	2.52	2.56	0.040000000000000036	l	s	197.7317027606333	195.1651482683486
32_F_CEC_phrase05.TextGrid	y	2.71	2.78	0.06999999999999984	s	d	297.0712243679064	375.2389585673768
32_F_CEC_phrase05.TextGrid	@	2.91	3.02	0.10999999999999988	d	l	197.06335430411514	200.2137271622694
32_F_CEC_phrase05.TextGrid	i	3.05	3.2	0.15000000000000036	l	d	246.29913169332545	228.9071515266359
32_F_CEC_phrase05.TextGrid	@	3.27	3.35	0.08000000000000007	d	t	180.17453085678696	187.06575142137206
32_F_CEC_phrase05.TextGrid	i	3.52	3.58	0.06000000000000005	t	p	288.65128117979157	317.840324628518
32_F_CEC_phrase05.TextGrid	a	3.66	3.72	0.06000000000000005	p	p	238.41566396065343	247.9297004138514
32_F_CEC_phrase05.TextGrid	a	3.83	3.94	0.10999999999999988	p	<p:>	185.24684172187793	188.4470565041405
32_F_CEC_phrase04.TextGrid	a	0.82	0.93	0.1100000000000001	<p:>	n	226.73757514327227	NA
32_F_CEC_phrase04.TextGrid	a	1.12	1.23	0.10999999999999988	m	R	221.7417714624666	223.99972291540456
32_F_CEC_phrase04.TextGrid	i	1.26	1.43	0.16999999999999993	R	e	250.40809494484225	241.10997847557778

Figure 3.3.2-1: extrait du fichier de sortie obtenu sur les voyelles du français pour les enregistrements avec microphone et carte son.

Cependant, le fichier tel quel ne nous a pas convenu, nous avons eu besoin d'ajouter quelques colonnes.

En premier lieu, nous avions fait le lien entre le fichier TextGrid dont sont extraites les analyses acoustiques et le fichier d'information du locuteur (« results_....txt ») grâce à un script PYTHON (ajoutcolonneresult_ENLIGNE.py, Annexe en ligne) puis avons finalement développé un script R car le programme Python prenait plusieurs heures à s'effectuer contre quelques secondes avec R. Dans le nom du fichier d'information du locuteur se trouvent des informations telles que l'âge, le sexe et l'identifiant du locuteur que nous avons intégré au fichier de sortie PRAAT grâce à un nouveau script R. Ce script R a donc créé les colonnes « nom du locuteur », « sexe du locuteur », « âge du locuteur » mais a aussi permis d'associer une classe d'âge à chaque locuteur, en fonction d'une table de correspondance établie dans un fichier texte externe (posttraitement_analyses_acoustiques_donneesenligne.R, Annexe en ligne).

4 RESULTATS

4.1 TRAITEMENT STATISTIQUES ET ANALYSE ACOUSTIQUE

4.1.1 Caractéristiques des locuteurs et des données

Grâce à de multiples analyses et bilans sur R, nous allons maintenant vous présenter les caractéristiques de notre corpus, qui se divise en deux sous corpus : d'un côté les données obtenues grâce à notre outil d'enregistrement en ligne, de l'autre côté les données enregistrées en présentiel à l'aide d'un microphone et d'une carte-son que nous appellerons par la suite « condition micro » ou condition « en présentiel ».

Pour les données enregistrées en présentiel, nous avons pu récupérer 429 enregistrements avec une moyenne de quinze enregistrements pour vingt-neuf locuteurs.

Pour les données enregistrées en ligne, nous n'avons gardé que les locuteurs pour lesquels nous avions au moins 12 enregistrements. Nous avons donc 1550 enregistrements pour 54 locuteurs.

Nous obtenons la répartition d'âges présentée en figure 4.1.1-1. Nous pouvons alors nous rendre compte qu'il y a un déséquilibre dans la répartition des locuteurs sur le plan de l'âge, pour les données en ligne comme pour les données en présentiel mais qu'il y a aussi un déséquilibre du point de vue de la taille des effectifs entre l'enregistrement en ligne et l'enregistrement avec le microphone et la carte son.

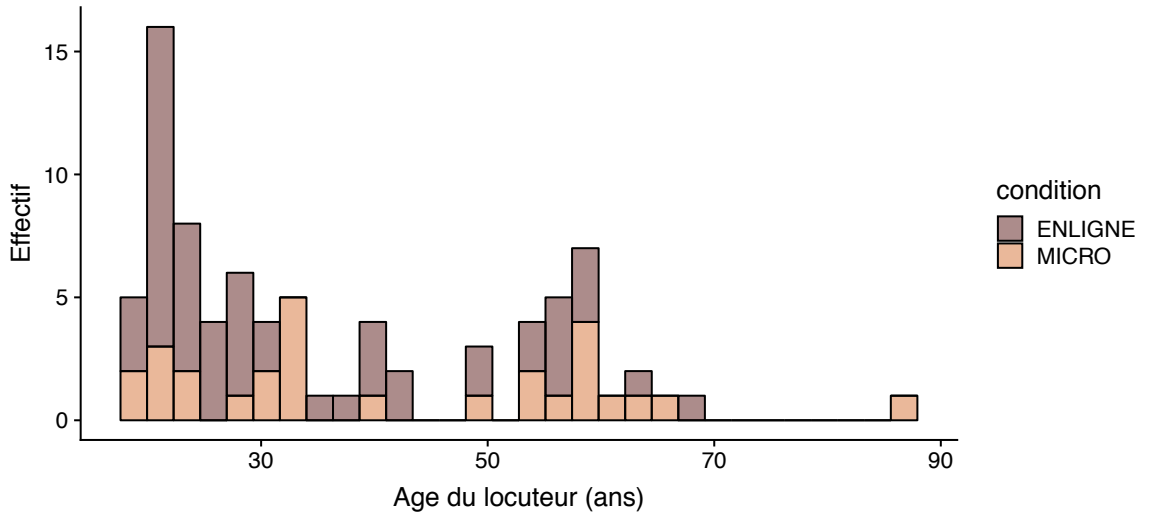


Figure 4.1.1-1: Répartition des locuteurs en fonction de leur âge pour les données en ligne et les données en présentiel.

Le déséquilibre en termes de classe d'âge est illustré par le tableau 4.1.1-1 Ainsi, nous pouvons nous rendre compte que la classe d'âge 20-29 et la classe d'âge 50-59 sont les plus largement représentées pour les données enregistrées en ligne respectivement 28 et 10 locuteurs, c'est-à-dire que plus de 50% des locuteurs qui se sont enregistrés en ligne sont de la classe d'âge 20-29 ans.

Tableau 4.1-1: Effectif des classes d'âges (en nombre de locuteur) en termes de phonèmes pour les données en présentiel et en ligne.

Classe d'âge	Effectif en présentiel	Effectif enligne
15-19	2	3
20-29	6	28
30-39	8	4
40-49	0	6
50-59	8	10
60-69	3	2
80-89	1	0

Si nous regardons d'encore plus près cette répartition en classe d'âge en ajoutant le critère du sexe du locuteur, nous nous rendons compte que la population masculine est largement moins bien représentée que la population féminine (figure 4.1.1-2).

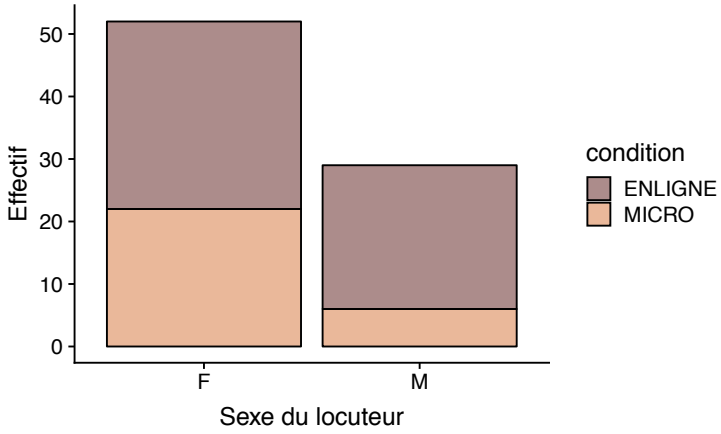


Figure 4.1.1-2: Répartition des effectifs (en nombre de locuteurs) en fonction du sexe pour les données en ligne et les données en présentiel.

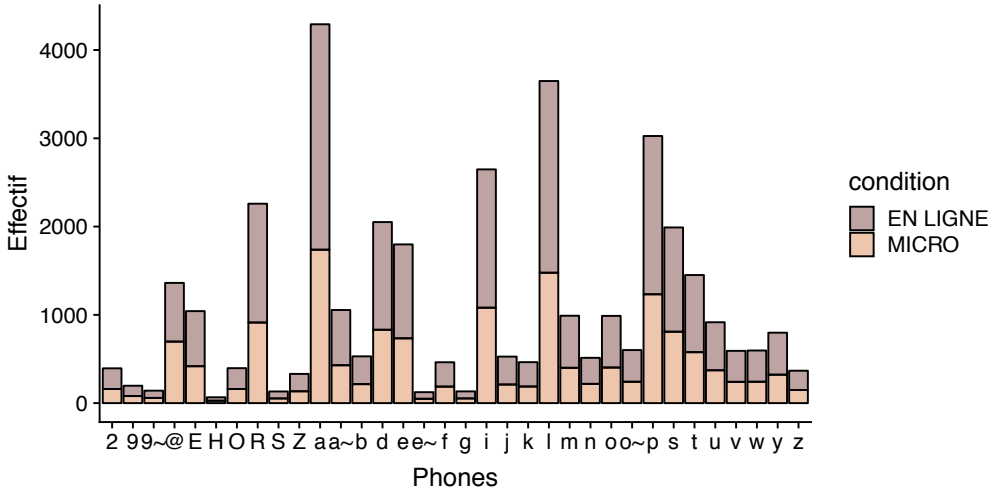
En effet, nous n'avons pas de représentants masculins des classes d'âges 15-19 et 80-89, alors que nous en avons chez les femmes. De plus, pour chaque classe des données en présentiel, nous possédons au minimum le double de locutrices que de locuteurs, nous observons une répartition un peu plus homogène pour les données récoltées en ligne (tableau 4.1.1-2).

Tableau 4.1-2 : Tableau comparatif des effectifs en fonction de la classe d'âge et du sexe du locuteur pour les données des deux sous-corpus.

Classe d'âge	Sexe	Condition « En présentiel »	Condition « En ligne »
15-19	F	2	3
20-29	F	5	16
30-39	F	7	1
40-49	F	0	3
50-59	F	5	6
60-69	F	2	1
80-89	F	1	0
15-19	M	0	0
20-29	M	1	12
30-39	M	1	3
40-49	M	0	3
50-59	M	3	4
60-69	M	1	1
80-89	M	0	0

Du point de vue de la répartition des sons dans le corpus nous avons également un fort déséquilibre (figure 4.1.1-3), les sons les plus représentés dont le /a/, /l/, /p/, /i/, /ʁ/. Nous obtenons, sans surprise, une distribution quasiment identique des phones pour les données en ligne et en présentiel, avec un effectif plus fort de 1,47 en moyenne pour les données en ligne.

Figure 4.1.1-3: Répartition des phones sur les deux sous-corpus.



Dans les transcriptions SAMPA, nous avions indiqué qu'il pouvait y avoir les deux variantes /ɛ̃/ et /œ̃/, mais en établissant un décompte de ces variantes en fonction de la phrase nous avons pu remarquer qu'un locuteur pouvait produire une variante ou une autre indépendamment du contexte et avons donc décidé de fusionner les deux phones pour ne garder que /ɛ̃/.

4.1.2 Variation intralocuteur

Nous avons donc récolté des données dans les deux conditions d'enregistrement pour un d'un panel de sujets. Nous avons donc récolté deux sous corpus de deux manières différentes, et nous souhaitons comparer les données recueillies en fonction de la condition d'enregistrement.

En effet, il est possible que pour un sujet, nous observions des résultats différents pour un même paramètre phonétique, en fonction de la condition d'enregistrement, même si la tâche demandée est une tâche de lecture. C'est ce que l'on qualifie de variation intralocuteur, en opposition avec la variation interlocuteur qui est attestée entre plusieurs individus distincts. Les facteurs qui peuvent jouer sur la variation intralocuteur pour notre panel de sujet enregistrés avec les deux conditions, sont l'environnement dans lequel les

sujets ont été enregistrés ou se sont enregistrés seuls, le fait d'avoir déjà rencontré ou non les énoncés (s'il s'agit de la première réalisation ou non de la tâche), de la distance laissée entre l'enregistrement des deux conditions, ainsi que tous les facteurs internes à la personne (humeur, fatigue, concentration…)

Nous avons pu mettre en parallèle les résultats des analyses acoustiques de Praat grâce à un code contrôle que nous avions attribué à notre panel de locuteurs au moyen de scripts R (traitement_donnees_comparaison.R, Annexe 3)

Nous pouvons ainsi comparer, pour un même locuteur, les valeurs d'un même paramètre tel que la fréquence fondamentale (F0), la durée des phones ou des phrases, les valeurs de formants pour les voyelles…

Pour la fréquence fondamentale, nous obtenons une différence moyenne d'environs 10.6 Hz (médiane à 9 Hz) avec une différence maximale de 23 Hz (troisième quartile à 17.2 Hz). Afin de se rendre compte du poids de cette différence, nous avons calculé le taux de variabilité entre conditions d'enregistrement pour chaque locuteur (tableau 4.1.2-1). Ce taux de variabilité est défini comme la différence entre la valeur obtenue en condition en ligne et en condition en présentiel, rapportée à la valeur en condition en présentiel considérée comme référence. Nous apprenons donc que le taux de variabilité moyen est d'environs 7% avec la médiane à 5% et une différence maximale à 21.6% (troisième quartile à 10%).

Tableau 4.1-3 : Valeur de F0 moyenne en Hertz par locuteur en fonction de la condition d'enregistrement.

Code contrôle	Age	Sexe	Valeurs F0 (Hz) Condition « En ligne » (Hz)	Valeurs F0 (Hz) Condition « En présentiel » (Hz)	Taux de variabilité (%)
MARINE	18	F	221	210	5
MOI	22	F	185	192	-4
PAU	22	F	221	214	3
TAT	22	F	183	182	1
CLA	23	F	205	186	10
MAN	23	F	211	214	-1
MAM	50	F	193	197	-2
VERO	56	F	197	211	-7
SYL	58	F	223	228	-2
ULR	29	M	129	106	22
HER	58	M	176	167	5
PAP	59	M	143	126	13
XAV	64	M	131	114	15

Nous pouvons également regarder la relation linéaire entre 2 variables, aussi appelée corrélation. La corrélation entre les valeurs de F0 en ligne et condition en présentiel pour les femmes est de 0.97, les variables sont donc fortement corrélées.

Cependant, nous pouvons noter que les valeurs sont plus hautes chez les hommes que chez les femmes qui présentent une moindre variabilité.

Pour les fréquences de formants F1, nous avons converti les valeurs initiales en Hertz vers des valeurs en Bark. L'échelle Bark est une échelle psycho-acoustique, basée sur une mesure subjective du son. C'est une échelle de fréquence sur laquelle chaque distance correspond à une distance perceptuelle égale et à partir de 500Hz, l'échelle de Bark devient de plus en plus linéaire.

Après conversion, nous obtenons une valeur de différence moyenne entre les deux conditions d'enregistrement d'environ 0.553 Bark soit 13% (médiane à 12.7%) avec une différence plus élevée pour les données en ligne que pour les données en présentiel, avec une différence maximale de 1.065 Bark, soit 29.7% (troisième quartile à 0.844 Bark). L'âge ne semble pas avoir d'incidence pour les données en ligne, mais pour les données en présentiel, si on regarde la corrélation entre l'évolution de l'âge et l'évolution de F1 en fonction du sexe, nous obtenons une corrélation de 0.590 pour les femmes et de 0.592 pour les hommes.

Les valeurs de formant F2 ont tendance à être légèrement plus élevées dans les enregistrements en présentiel pour les hommes et est généralement plus élevée pour les données enregistrées en ligne pour les femmes (tableau 4.1.2-2). Les différences n'ont pas l'air d'être liées à l'âge. Nous pouvons observer une différence moyenne de 4% (médiane à 1.3%) avec une valeur maximale de 16% (troisième quartile à 3%).

Code contrôle	Valeurs de F2 condition « En ligne » (Bark)	Valeurs de F2 condition « En présentiel » (Bark)	Taux de variabilité (%)
MARINE	11.4	11.76	-3.1
MOI	11.99	11.85	1.2
PAU	12.13	12.11	0.2
TAT	9.82	11.8	-16.8
CLA	12.07	12.14	-0.6
MAN	12.02	10.92	10.1
MAM	11.86	11.75	0.9
VERO	12.05	11.82	1.9
SYL	10.07	11.41	-11.7
ULR	11.37	10.96	3.7
HER	11.05	10.9	1.4

PAP	11.15	11.22	-0.6
XAV	10.81	10.7	1

Tableau 4.1-4: Valeurs de F2 moyennes en Bark et taux de variabilité par locuteur en fonction de la condition d'enregistrement (classé par sexe [MARINE-SYL = F] et par âge croissant).

Les valeurs de formants F3 ont tendance à être plus élevées pour les enregistrements condition en présentiel que pour les enregistrements en ligne pour les femmes et inversement pour les hommes. La différence moyenne est de 2.2% (médiane à 1.88%), la différence maximale est de 8.5% (troisième quartile à 2.7%).

Pour la différence de longueur des phrases, nous observons que pour un même locuteur, les énoncés enregistrés en ligne sont généralement un peu plus longs que les énoncés enregistrés avec le microphone et la carte son. Cette différence est en moyenne de 0.378 secondes soit 9% (médiane à 0.256 secondes, soit 5.9%) avec une différence maximale de 1.19 secondes soit 27.8% (troisième quartile à 0.54 secondes soit 12%).

Pour les données de durée de phrase, nous obtenons des valeurs de corrélation élevées, pour les données en ligne des hommes (0.805) et des femmes (0.851) et pour les données en présentiel des femmes (0.787) (tableau 4.1.2-3).

Tableau 4.1-5 : Valeur de durée moyenne des énoncés en secondes par locuteur en fonction de la condition d'enregistrement (classé par sexe [MARINE-SYL = F], et par âge croissant).

code_controle	En ligne	En présentiel	difference
MARINE	3.28	3.28	0
MOI	3.9	3.81	2.4
PAU	3.93	3.39	15.9
TAT	3.06	3.26	-6.1
CLA	3.59	3.39	5.9
MAN	3.8	4.35	-12.6
MAM	4.16	4.1	1.5

VERO	4.56	4.31	5.8
SYL	5.41	5.15	5
ULR	3.75	4.22	-11.1
HER	5.71	5.59	2.1
PAP	5.48	4.29	27.7
XAV	4.77	3.8	25.5

En faisant le rapport entre la différence entre conditions pour chaque locuteur (variation intra-locuteur) et l'étendue des valeurs pour l'ensemble des locuteurs (variation inter-locuteurs). Nous obtenons une variation de 7% pour la durée, de 6% pour la fréquence fondamentale et de 5% pour le formant F2.

Nous pouvons donc noter qu'il y a de la variation intralocuteur, mais celle-ci reste minime et évolue de la même manière selon l'âge et le sexe, elle est, de plus, moindre que la variation interlocuteur.

4.1.3 Cout de l'enregistrement non-contrôlé par rapport à l'enregistrement contrôlé, calcul du rapport signal sur bruit (signal-to-noise ratio, SNR)

4.1.3.1 Comparaison des SNR des données des locuteurs enregistres en ligne et en présentiel

Le calcul du SNR (Signal to Noise Ratio) permet de connaitre la qualité de l'enregistrement en termes de bruit de fond. Sur nos données, nous le définissons comme la soustraction de la valeur d'énergie moyenne en dB dans les voyelles et de l'énergie moyenne en dB dans les silences. Ainsi, plus le SNR est élevé, moins le bruit de fond est fort et inversement, un SNR faible montre un signal très bruité.

Nous avons donc voulu comparer les SNR des données enregistrées en ligne et le SNR des données enregistrées avec la carte son. Nous nous attendons à avoir un SNR

meilleur pour les enregistrements effectués en présentiel, étant donné qu'il s'agit du contexte contrôlé.

Contre toute attente, nous avons observé des SNR nettement meilleurs pour les enregistrements en ligne que pour les enregistrements en présentiel (tableau 4.1.3-4).

Code contrôle	SNR condition « enligne »	SNR condition « en présentiel »
CLA	221	33
HER	75	37
MAM	50	32
MAN	57	37
MARINE	64	38
MOI	56	38
PAP	78	
PAU	176	
SYL	65	37
TAT	53	31
ULR	65	37
VERO	147	28
XAV	136	35

Tableau 4.1-6 : Comparaison des SNR pour les données en ligne et condition en présentiel des locuteurs ayant enregistré des deux manières.

En regardant de plus près les résultats, nous pouvons nous rendre compte que pour certains enregistrements en ligne, nous avons des valeurs moyennes d'intensité en dB négatives. En regardant dans lesdits enregistrements, nous observons que sur les temps de pause, la courbe est totalement lisse, comme s'il n'y avait aucun son produit en fond, ce qui est impossible même dans les meilleures conditions d'enregistrement possibles avec le meilleur matériel qui soit. Il doit donc s'agir d'un post-traitement qui, en dessous d'un certain seuil d'énergie acoustique, efface totalement le bruit de fond. Nous ne pourrons donc pas

utiliser le SNR comme indicateur de qualité d'enregistrement pour les données recueillies en ligne de ce fait. Nous pouvons cependant regarder si, en fonction de la classe d'âge nous observons des différences notables dans les valeurs de SNR.

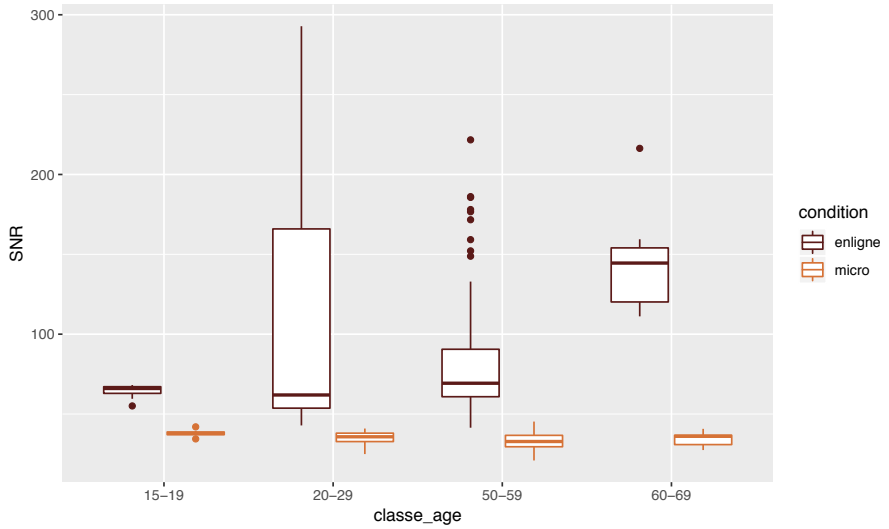


Tableau 4.1-7 : Comparaison des valeurs de SNR en fonction de la classe d'âge pour chaque condition d'enregistrement (condition « micro » = « en présentiel »).

Nous pouvons alors observer en figure 4.1.3-2 que les valeurs de SNR pour les données en présentiel ont tendance à diminuer très légèrement alors que, pour les données en ligne, bien qu'elles soient biaisées du fait que le bruit ait été post-traité pour certains enregistrements, nous observons une très nette augmentation du SNR pour la classe d'âge 60-69 ans, et des valeurs très hétérogènes pour la classe 20-29 ans, nous obtenons des valeurs de corrélation entre le SNR et l'âge de -0.26 pour les données en condition « présentiel » et de -0.0013 pour les données en condition « enligne ». Si nous ne pouvons pas exploiter de manières brutes ces valeurs de SNR pour les données en ligne, nous pouvons tout de même émettre des hypothèses sur cette évolution des valeurs de SNR.

En effet, nous pourrions supposer que cette augmentation du SNR révèle une différence d'utilisation et de posture du locuteur par rapport au microphone en fonction de l'âge, différence qui n'aurait pas été perçue pour les enregistrements avec microphone et carte-son car le microphone était placé sur la tête du locuteur et ajusté par l'expérimentatrice de la même manière pour chaque locuteur. Une autre hypothèse serait que les différences de variations des valeurs de SNR reposeraient sur la qualité de la segmentation automatique et donc seulement sur les petites portions de son qui ont été mal découpées en étant considérées à tort comme des pauses silencieuses.

4.1.3.2 Influence du SNR sur la classification

Afin de savoir si ces valeurs de SNR ont une influence sur la classification, nous avons effectué une classification seulement sur ces valeurs. Si nous obtenions de bons résultats de classification, nous en aurions conclu que l'algorithme classait en fonction du SNR plutôt qu'en fonction des autres paramètres acoustiques.

Avec l'algorithme J48, que nous utiliserons plus tard pour nos classifications avec les MFCC et les paramètres phonétiques, nous obtenons un taux d'instances correctement classées de 52.8 % (204 instances bien classées sur 386), avec une précision, rappel et F-mesure respectivement de 0.516, 0.528 et 0.520, ce qui équivaut au hasard défini par l'algorithme ZeroR (53%).

4.1.3.3 Influence du SNR sur la totalité des données enregistrées en ligne

Nous allons maintenant vous présenter les résultats de cette analyse appliquée à la totalité de nos données enregistrées en ligne.

Regardons d'abord la répartition des valeurs de SNR en fonction de la classe d'âge pour toutes nos données en ligne.

Nous observons le même type de valeurs, avec parfois des SNR très élevés dues au post-traitement du microphone utilisé par le locuteur. La figure 4.1.3-1 permet de voir, sur le graphe de gauche, qu'il y a des données post-traitées pour tous les âges et le graphe de droite permet de voir s'il y a une tendance globale d'évolution du SNR. Les classes d'âge 15-19 et 60-69 ans ne sont composées que d'enregistrements ayant fait l'objet d'un post-traitement, c'est pourquoi nous observons une montée abrupte du SNR moyen pour ces deux classes. Pour les autres classes, pour lesquelles nous observons des valeurs moyennes de SNR issues à la fois de données post-traitées et non post-traitées, nous pouvons voir que la valeur de SNR moyen par classe a tendance à baisser légèrement.

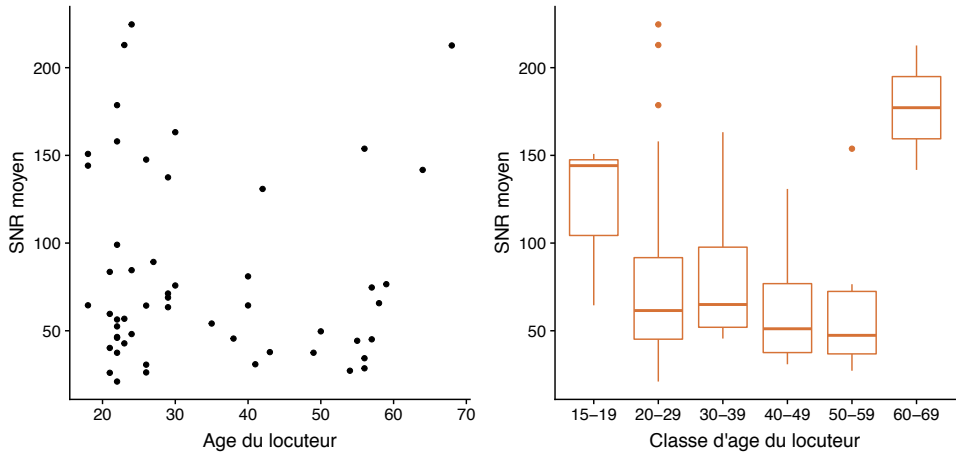


Figure 4.1.3-1 Comparaison des valeurs de SNR moyen en fonction de l'âge (à gauche) ou de la classe d'âge (à droite).

En appliquant le même algorithme aux valeurs de SNR pour toutes nos données en ligne, nous obtenons un taux d'instances bien classées de 50.58 % (392 instances bien classées contre 383 mal classées), avec des valeurs de précision, rappel et F-mesure respectivement égales à 0.385, 0.506 et 0.388. Ces chiffres montrent que la classification est très probablement le résultat du hasard. Nous pouvons donc considérer que la classification ne se fera pas en fonction du SNR et nous pourrons donc utiliser des paramétrisation du type MFCC.

4.1.4 ANALYSE DES DONNEES EN LIGNE

4.1.4.1 La durée ou débit articulatoire

Nous avons étudié la durée globale des phones pour les deux sous-corpus (figure 4.1.4-1). Ainsi, nous pouvons remarquer que la longueur du phone a une tendance globale à s'allonger avec l'âge et que la condition d'enregistrement semble avoir une très faible incidence sur la longueur de la réalisation du phone.

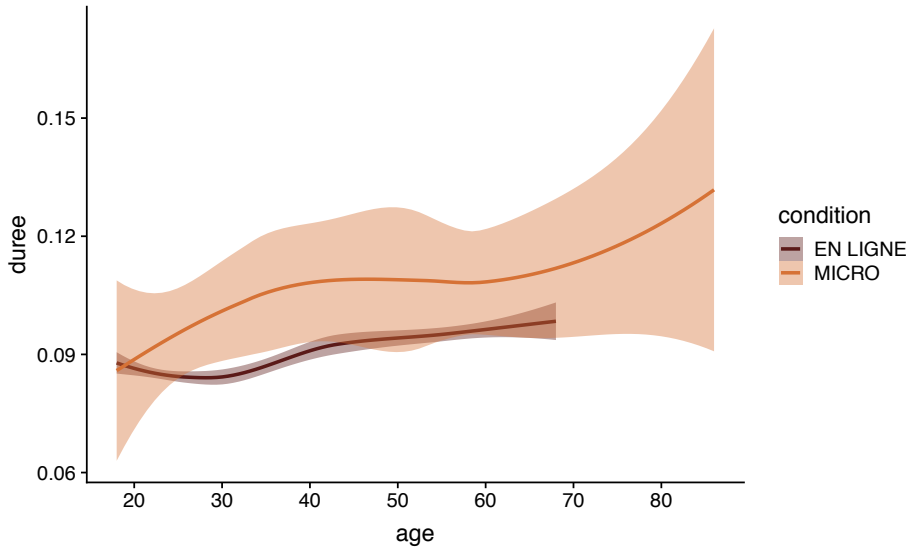


Figure 4.1.4-1: Évolution de la longueur du phone en fonction de l'âge du locuteur pour les deux sous-corpus, tout sexe confondu (l'enveloppe autour des courbes correspond à la variabilité pour les différents âges considérés).

Mais il n'est pas suffisant de s'arrêter à une tendance globale, en effet en fonction du type de phones (voyelle ou consonne), du point d'articulation, du mode d'articulation ou encore du voisement, nous pouvons supposer qu'il y aura des différences de variation.

En nous intéressant plus spécifiquement aux données en ligne, nous pouvons regarder l'évolution de la durée du phone en fonction du phone. Nous avons tenté de voir au niveau des consonnes s'il y avait une différence entre consonnes fricatives et consonnes occlusives en fonction de leur voisement et du sexe du locuteur.

Ainsi pour les fricatives (figure 4.1.4-2), nous pouvons observer que la durée est significativement plus longue pour les sourdes que pour les sonores, tout sexe confondu,

et que la variation en fonction de l'âge est nettement plus forte pour les fricatives sourdes. Nous pouvons noter également l'accourcissement de la durée de la fricative sourde a lieu plus tôt chez les hommes que chez les femmes.

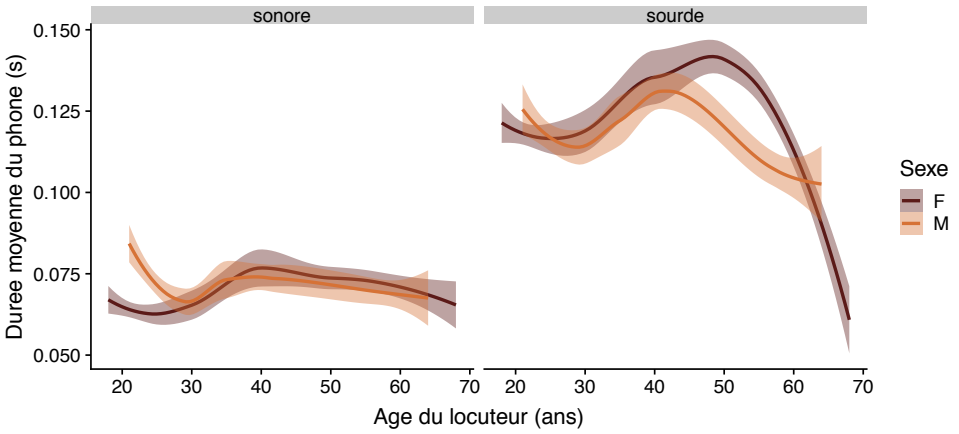


Figure 4.1.4-2 : comparaison des durées moyennes des fricatives sonores et sourdes en fonction du sexe du locuteur pour les données enregistrées en ligne.

Nous pouvons également regarder l'évolution de la durée des phones de type voyelles (figure 4.1.4-3). Nous pouvons voir que la durée varie plus ou moins en fonction de la phrase. La phrase pour laquelle la durée de la voyelle connait le plus de variation est la première phrase du corpus, il s'agit de la plus courte. Pour les autres phrases nous pouvons observer que la durée moyenne de la voyelle a tendance à augmenter.

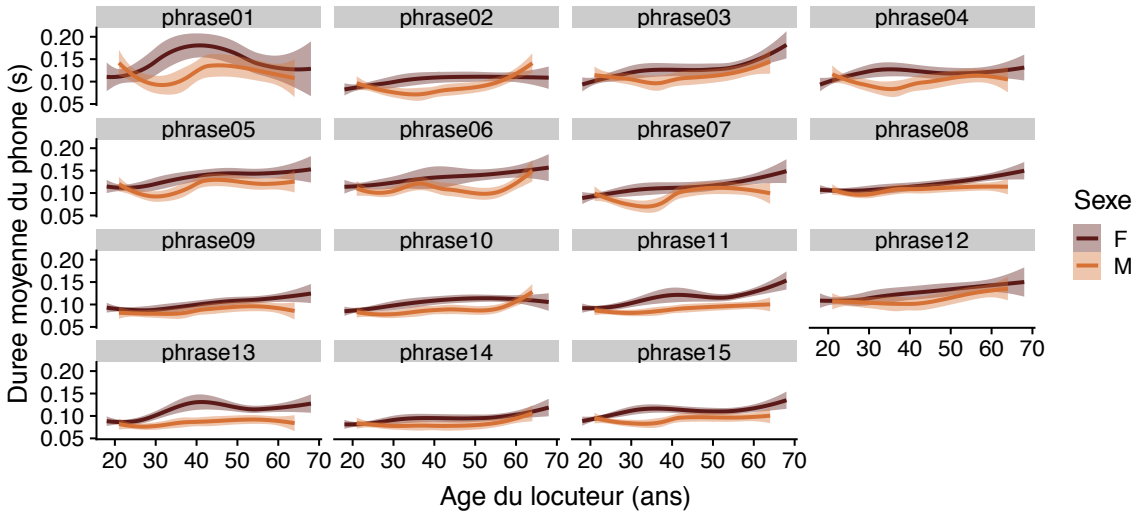


Figure 4.1.4-3 : évolution de la durée de la voyelle, en fonction de l'âge et du sexe du locuteur, par phrase, pour les données enregistrées en ligne.

En regardant la répercussion de ces allongements et raccourcissements de temps de production des phones au niveau du débit articulatoire, c'est-à-dire la durée de la phrase sans prendre les pauses en compte, nous pouvons constater que la longueur des phrases a tendance à très faiblement s'allonger, nous avons pu mettre en avant une corrélation moyenne par phrase de 0.4984996 pour les femmes et de 0.3297142 pour les hommes entre les variables de l'âge et de la durée de la phrase.

4.1.4.2 Le débit

Nous pouvons alors nous demander si le débit évolue avec l'âge pour les données de notre corpus. Nous pouvons observer que la durée totale de l'énoncé à tendance à s'allonger avec l'âge, aussi bien chez les hommes que chez les femmes (figure 4.1.4-4), en notant cependant que les hommes produisent des énoncés légèrement plus courts. Cela signifie donc qu'avec l'âge le débit baisse.

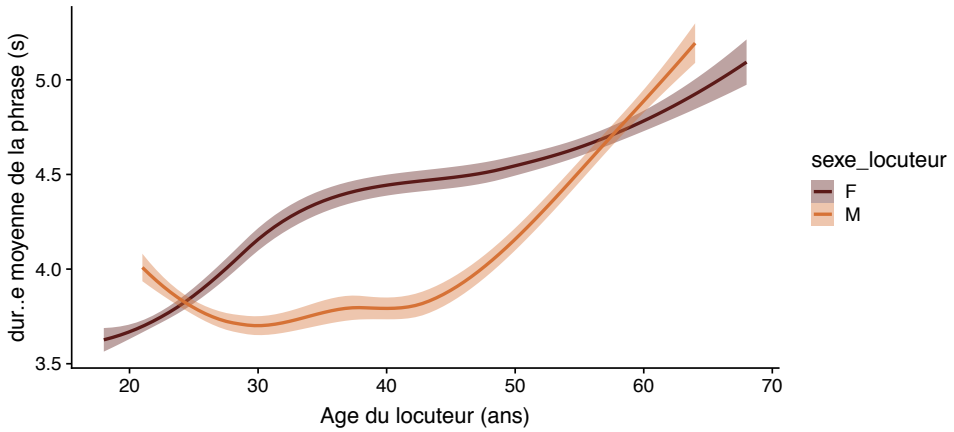


Figure 4.1.4-4 : Comparaison de l'évolution de la durée totale de l'énoncé (débit) en fonction de l'âge et du sexe du locuteur.

Il est également intéressant de regarder l'évolution du débit au niveau de la phrase. En effet, nous nous rendons compte sur la (figure 4.1.4-5.) que pour les quatre premières

phrases du corpus, la durée moyenne de la phrase en comptant les pause a tendance a augmenter de manière plutôt douce, la diminution du débit est très faible, alors qu'à partir de la phrase cinq et jusqu'à la phrase quinze, nous avons une augmentation plus nette de la durée de la phrase (jusqu'à deux secondes) et donc une diminution du débit plus importante. La « phrase01 » correspond à la phrase la plus courte du corpus avec quatre mots, alors que les phrases huit et quinze, par exemple, sont les phrases les plus longues, respectivement dix-huit et quinze mots. Rappelons que pour les données enregistrées en ligne, la numérotation des phrases est indépendante de l'ordre dans lequel elles ont été produites par les locuteurs, les phrases ayant été présentées en ordre aléatoire lors de l'enregistrement.

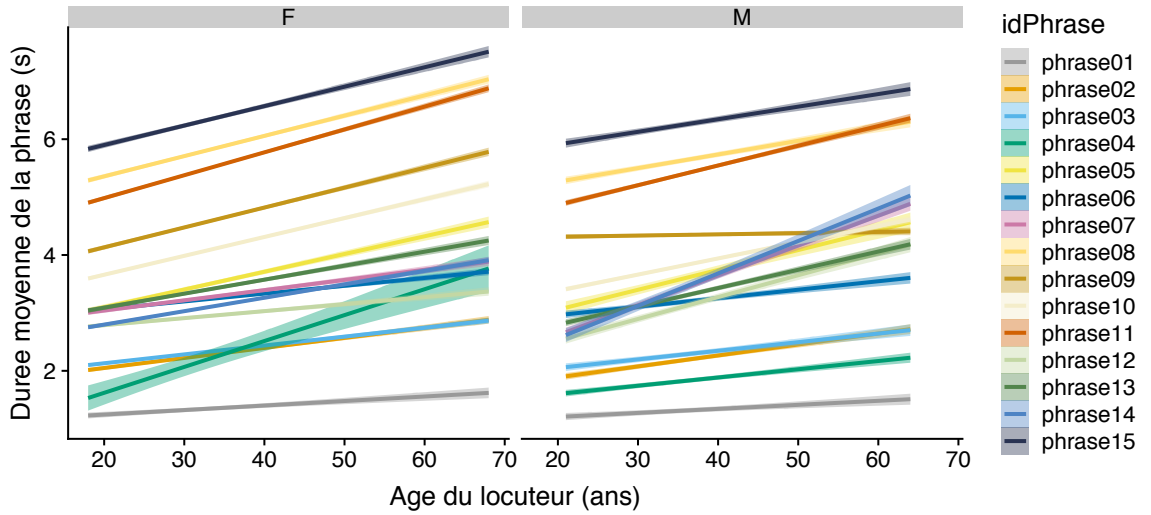


Figure 4.1.4-5 : Évolution de la durée de l'énoncé en fonction de l'âge et du sexe du locuteur, par phrase, pour les données enregistrées en ligne.

Ici, nous avons utilisé la méthode « lm » de geom_smooth, car elle renvoie un résultat lissé et linéaire et nous cherchions à rendre compte d'une tendance d'évolution, et comme les données en présentiel pour les hommes sont insuffisantes, il n'était pas judicieux de les représenter.

4.1.4.3 *La fréquence fondamentale F0*

Nous vous avons précédemment présenté le résultat d'études sur l'évolution de la fréquence fondamentale avec le vieillissement de la voix, nous allons, alors, nous intéresser aux résultats de l'extraction de Praat pour la F0. Nous comparerons toujours les données enregistrées avec le microphone et la carte son avec les données enregistrées en ligne, tout en séparant les locuteurs des locutrices.

Ainsi, comme constatable sur les figures 4.1.4-6 et 4.1.4-7, nous observons que les hommes ont une fréquence fondamentale (F0) plus faible que les femmes.

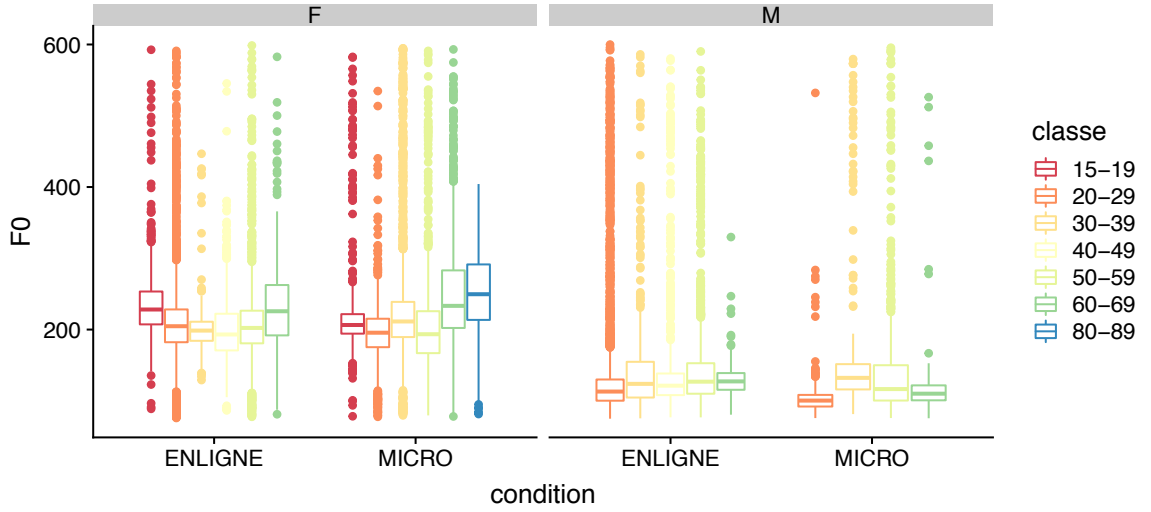


Figure 4.1.4-6: F0 moyenne (tous phones confondus) pour les données des deux sous-corpus par classe d'âge (condition « MICRO » = présentiel).

Les données enregistrées en présentiel pour les hommes sont insuffisantes mais en se concentrant sur les femmes, nous observons que l'évolution de la fréquence fondamentale est similaire pour les deux conditions. Pour les données enregistrées en ligne, nous observons une variation plus forte de la F0 pour les femmes que pour les hommes.

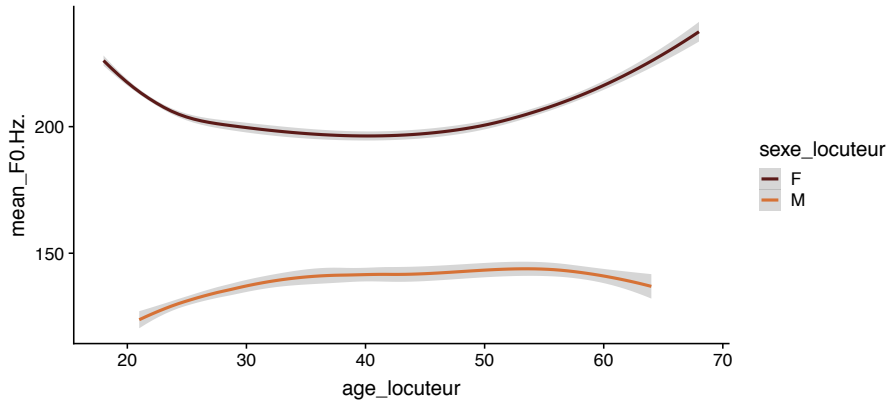


Figure 4.1.4-7 : Fréquence fondamentale moyenne (F0) en fonction de l'âge et du sexe du locuteur pour les enregistrements en ligne.

4.1.4.4 Les formants

Nous portons notre analyse formantique sur les voyelles de notre corpus en excluant d'emblée les voyelles nasales, pour lesquelles la détection des formants n'est pas fiable.

Les valeurs de formants F1 des voyelles, varient très peu d'une classe d'âge à une autre, sauf pour les voyelles /a/ et /ɛ/, dont les valeurs ont tendance à varier à partir de 50 ans. Les valeurs de formants F1 augmentent chez les femmes et diminuent chez les hommes (figure 4.1.4-8).

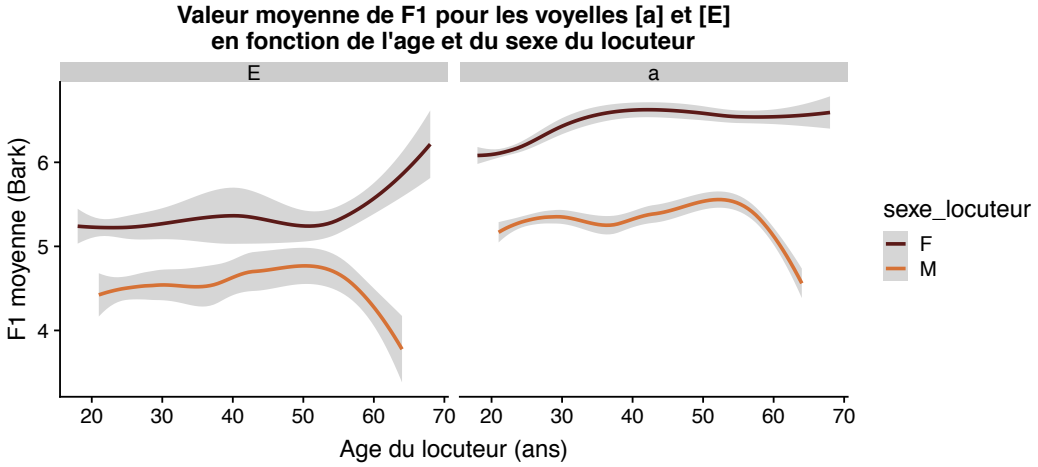


Figure 4.1.4-8 : Valeurs moyennes des fréquences de formants F1 (Bark) en fonction de l'âge et du sexe du locuteur pour les voyelles /a/ et /ɛ/.

Si les valeurs moyennes de F1 ont tendance à augmenter avec l'âge chez les femmes pour certaines voyelles, les valeurs de F2 ont, elles, tendance à diminuer avec l'âge de manière très abrupte pour le /a/, le /œ/ et le /o/ à partir de cinquante ans. Cette baisse abrupte de la valeur moyenne de formant F2 est également observable chez les hommes (figure 4.1.4-9).

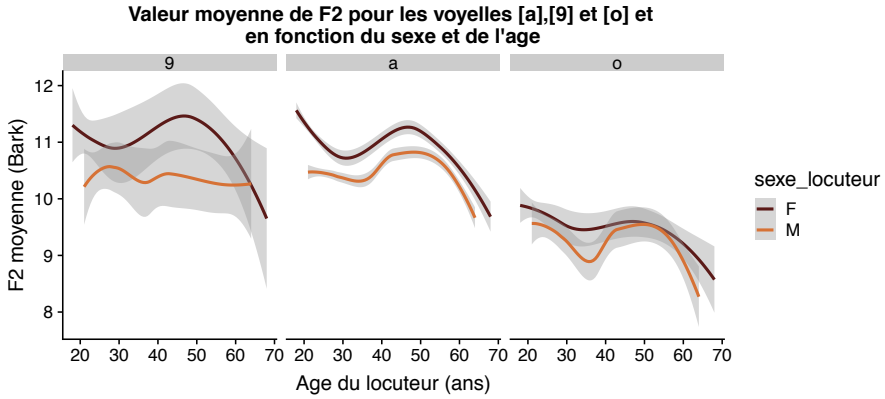


Figure 4.1.4-9: Valeurs moyennes de F2 (Bark) pour les voyelles nasales /a/, /o/ et /œ/ en fonction de l'âge et du sexe des locuteurs enregistrés en ligne.

Pour les valeurs moyennes de F3, nous pouvons observer des variations plus fortes sur les voyelles /a/, /o/, /ɔ/, /u/ (figure 4.1.4-10) chez les femmes et une légère augmentation des valeurs de F2 pour les voyelles /ɔ/ et /y/ chez les hommes.

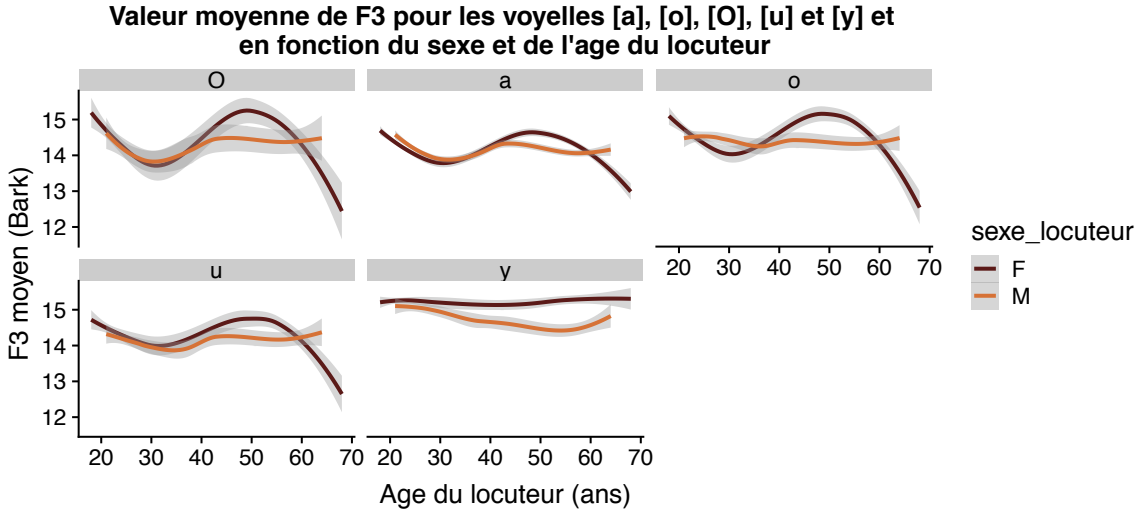


Figure 4.1.4-10 : Valeur moyenne de F3 pour les voyelles /a/, /o/, /ɔ/, /u/ et /y/ en fonction de l'âge et du sexe des locuteurs enregistrés en ligne.

4.1.4.5 *Le ZCR*

Le Zero-Crossing Rate (ZCR) pouvant être considéré comme une mesure du degré d'apériodicité dans la parole, nous avons décidé de voir s'il était impacté par le vieillissement du locuteur. Nous avons donc dans un premier temps concentré notre analyse sur les sons voisés du corpus, et plus particulièrement les voyelles car elles sont par nature périodique, puis dans un second temps, nous avons analysé les valeurs de taux de passage par zéro des consonnes occlusives sourdes et fricatives sourdes qui sont par nature apériodique.

4.1.4.5.1 Le ZCR sur les phones de nature périodique

Nous nous sommes donc demandés si le taux de passage par zéro (ZCR) évoluait en fonction de l'âge pour les sons voisés du corpus. Nous avons regardé la variation du ZCR au niveau des voyelles, et nous avons pu mettre en avant une variation plus forte pour les voyelles /ɛ/, /i/, /y/ dont les valeurs moyennes baissent fortement pour les femmes comme pour les hommes à partir de quarante ou cinquante ans. Ce résultat peut sembler contre-intuitif, car il suggère que les sons voisés seraient moins périodiques pour les locuteurs plus jeunes que pour les locuteurs plus âgés. Pour les autres voyelles, les valeurs de ZCR restent stables (figure 4.1.4-11).

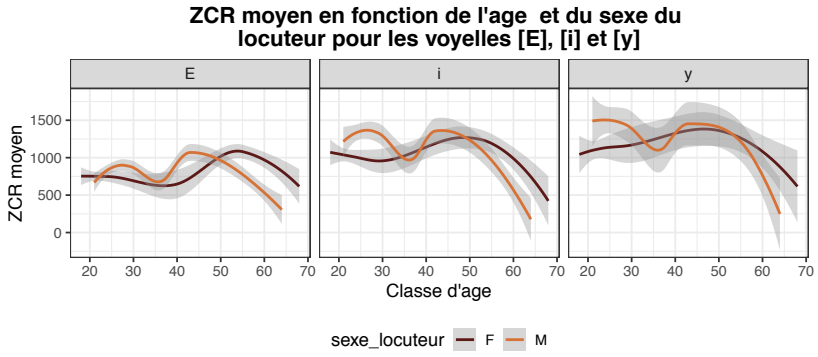


Figure 4.1.4-11: Valeurs de ZCR moyennes pour les voyelles /ɛ/, /i/ et /y/ en fonction de l'âge et du sexe du locuteur.

Nous nous sommes également intéressées aux consonnes voisées du corpus et ce sont les consonnes fricatives sonores /z/ et /ʒ/ qui présentent de plus fortes variations (figure 4.1.4-12). en effet, pour le /ʒ/, à partir de quarante ans, le ZCR connait une forte baisse, chez les femmes cette baisse est plus légère et plus tardive (cinquante-cinq ans). Pour le /z/, nous observons la même baisse que pour le /ʒ/ chez les hommes et la baisse chez les femmes commence plus tôt (cinquante ans) et est plus abrupte.

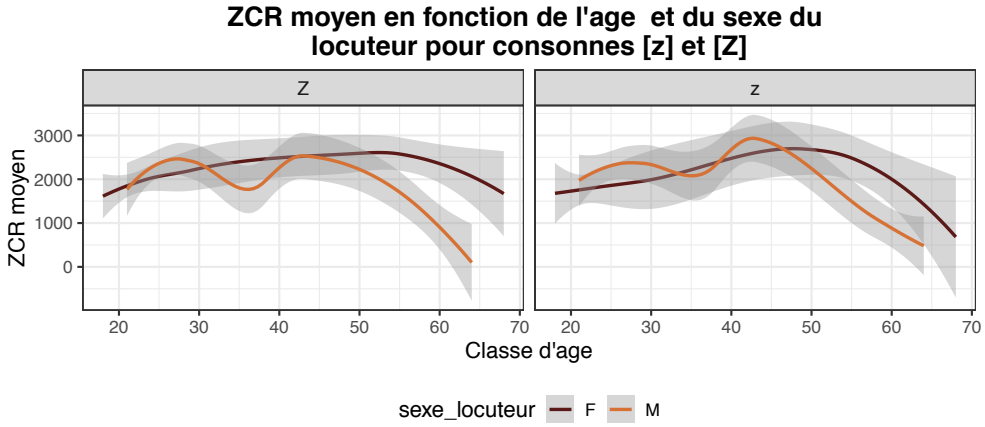


Figure 4.1.4-12 : Valeurs moyennes du ZCR pour les consonnes /z/ et /ʒ/ en fonction de l'âge et du sexe du locuteur enregistré en ligne.

4.1.4.5.2 Le ZCR sur les phones de nature apériodique

La variation du ZCR moyen des phones de nature apériodique semble similaire à la variation des phones de nature périodique, c'est-à-dire une première hausse du ZCR chez lès hommes à partir de quarante ans, puis une forte baisse à partir de cinquante ans et pour les femmes, un ZCR moyen qui connait une première baisse jusqu'à trente ans, puis une stabilisation jusqu'à et enfin une forte baisse à partir de cinquante ans.

Nous pouvons remarquer que la valeur moyenne de ZCR pour les fricatives sourdes est supérieure à celle des occlusives sourdes ce qui n'est pas surprenant puisque la friction est maintenue plus longuement pour les fricatives que pour les occlusives (figure 4.1.4-13).

De nouveau, ces résultats sont surprenant puisque cela signifie que les sons apériodiques ont tendance à devenir périodique avec l'âge.

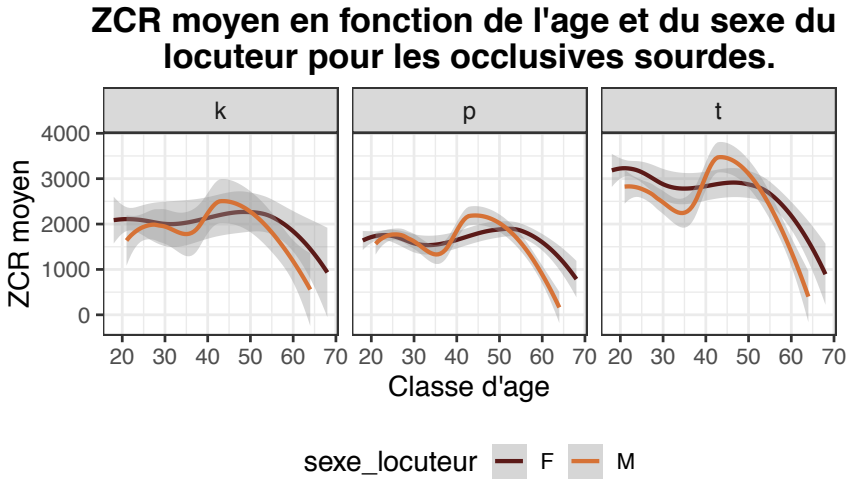


Figure 4.1.4-13: Taux de passage par zéro (ZCR) moyen en fonction de la classe d'âge par consonne occlusive sourde, pour les données en ligne.

Les taux de passages par zéro moyens semblent évoluer de la même manière pour chaque consonne occlusive sourde et le /ʃ/, quelque soit la condition d'enregistrement, on observe une augmentation du ZCR moyen entre la classe 15-19 et la classe 20-29, puis une baisse du ZCR moyen entre les classes 20-29 et 30-39, puis une nouvelle hausse pour la classe 50-59, une dernière baisse du ZCR pour la classe 60-69 pour finir sur une hausse forte pour la classe 80-89 (figure 4.1.4-14).

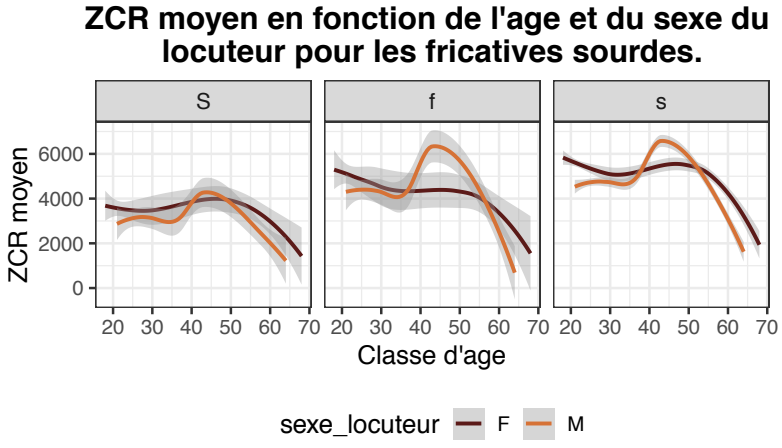


Figure 4.1.4-14 : Taux de passage par zéro (ZCR) moyen en fonction de la classe d'âge par consonne fricative sourde, pour les données en ligne.

Pour /f/ et /s/, nous obtenons respectivement un ZCR moyen bas (espace entre le premier et le troisième quartile est très large mais la médiane le situe très bas) et un ZCR

haut pour la classe 15-19 ans puis une baisse pour les classes suivante jusqu'à obtenir la valeur de ZCR moyen la plus haute pour la classe 80-89 ans.

La classe 15-19 ans est la classe pour laquelle nous avons le plus de variation intercondition.

4.1.4.6 Le Centre de gravité spectrale (CGS)

Nous pouvons également nous intéresser au centre de gravité spectral qui est un indicateur du lieu d'articulation dans le conduit vocal, principalement pour les fricatives. Effectivement, nous avons pu observer des valeurs de CGS exploitables pour les fricatives sourdes alvéolaires et labio-dentales (figure 4.1.4-15). Pour le /ʃ/, nous remarquons un centre de gravité spectrale moyen stable jusqu'à cinquante ans chez les femmes puis une baisse alors que pour les hommes, les valeurs moyennes du centre de gravité spectrale commencent par monter jusque quarante ans, pour ensuite baisser également de manière plus abrupte. Pour le /s/ et le /f/, nous observons des variations similaires, chez les femmes, il y a une première baisse du CGS entre quinze et trente ans, puis le CGS se stabilise jusqu'à cinquante ans pour de nouveau baisser assez fortement, alors que pour les hommes, les valeurs moyennes de CGS sont stables de vingt à trente ans puis connaissent une très forte hausse entre trente-cinq et quarante-cinq ans pour enfin connaitre une très forte baisse à partir de cinquante ans.

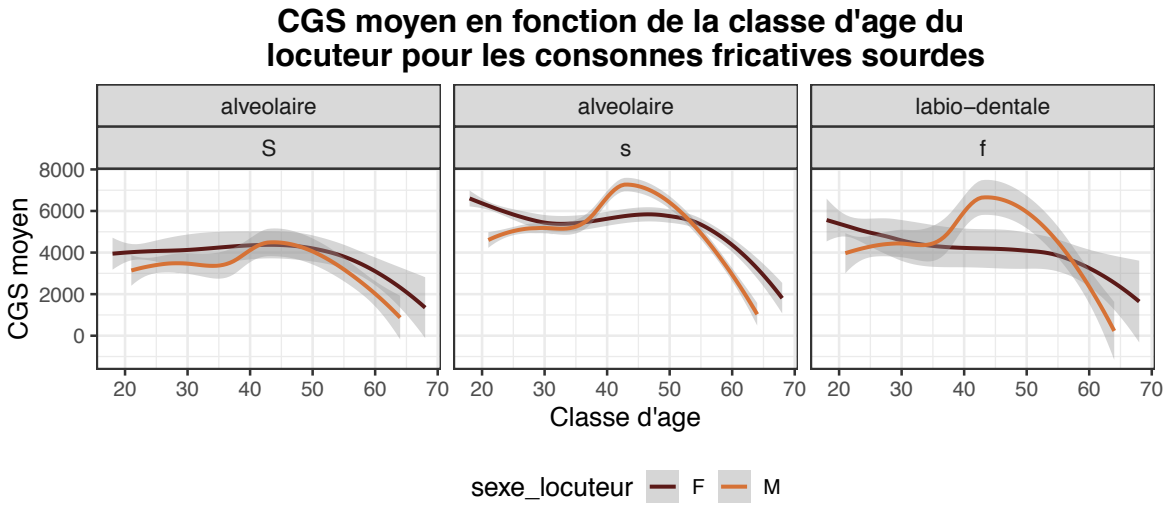


Figure 4.1.4-15 : CGS moyen en fonction de la classe d'âge du locuteur pour les consonnes fricatives sourdes alvéolaires et labio-dentales des données enregistrées en ligne.

4.2 DATA MINING

4.2.1 CHOIX DU TYPE DE CLASSIFICATION

Nous nous situons dans le domaine de la classification supervisée. En effet, nous avons préétabli des classes et demandons à Weka de classifier en fonction de ces classes.

4.2.2 CHOIX DES DESCRIPTEURS

Nous choisissons d'opposer les MFCC et les paramètres phonétiques afin de déterminer si on obtient de meilleures performances en utilisant des paramètres extraits automatiquement, sans avoir la connaissance de ce qu'ils mesurent, ou en ayant recours à des paramètres phonétiques plus directement interprétables, qui permettent donc de mieux comprendre quels sont les dimensions qui contribuent plus ou moins aux différences acoustiques entre classes d'âges dans les échantillons produits par les locuteurs. Nous utiliserons donc les MFCC comme une *baseline*, une référence à surpasser.

4.2.2.1 MFCC

Nous avons extrait les coefficients MFCC grâce à un script Python (Annexe 4), qui utilise la bibliothèque python_speech_features[3] et nous avons choisi d'extraire dix-neuf coefficients (Meigner and Rouvier 2012) (Kahn 2011) sur des fenêtres de 0.025 secondes de longueur, espacées de 0.01 secondes. Ce script fournissait un jeu de MFCC par phrase et par locuteur, mais nous souhaitions ne conserver qu'un seul vecteur par phrases de chaque locuteur, nous avons donc post-traité la sortie de Python avec R (« MFCC_group.R » Annexe en ligne) pour obtenir la moyenne et l'écart-type de chaque coefficient pour chacune des phrases, et avons reliés chaque ligne à la classe d'âge correspondant. Les dérivées de chaque coefficient (souvent appelées *delta* dans les applications en traitement automatique de la parole) afin de rendre compte de l'évolution à court terme du signal de parole.

Nous n'avons pas souhaité extraire les MFCC avec Praat, bien que cela soit facilement mis en place car il a déjà été constaté que les valeurs pouvaient différer des autres extractions de MFCC.

4.2.2.2 Paramètres phonétiques

Pour les paramètres phonétiques, nous avons donc retenu les valeurs de moyenne et d'écart-type de la fréquence fondamentale pour les voyelles, les valeurs moyennes des formants F1, F2 et F3 pour le /a/ ainsi que leurs valeurs d'écart-type, car il s'agit de la voyelle la plus fréquente de notre corpus, La valeur moyenne du CGS et son écart-type pour les consonnes occlusives sourdes, les valeurs de moyenne et d'écart-type du ZCR pour les voyelles /i/ et /y/, les consonnes fricatives sourdes /f/ et /s/, et les consonnes

[3] https://python-speech-features.readthedocs.io/en/latest/

fricatives sonores /z/ et /ʒ/, la durée moyenne du phone pour les fricatives d'un côté et pour les voyelles de l'autre et enfin la durée totale de la phrase avec les pauses (débit).

4.2.3 CLASSIFICATION

Nous allons vous présenter les résultats de classifications par différents algorithmes pour les données enregistrées en ligne, en comparant, la classification pour laquelle les MFCC sont les descripteurs et la classification pour laquelle les descripteurs sont des paramètres phonétiques.

Nous vous présenterons les taux d'instances bien classées, ainsi que certaines des matrices de confusion et les valeurs de précision, rappel et F-mesure pour chaque classification.

La matrice de confusion présente les taux d'instances classées par classes prédites en fonction de la classe de référence en pourcentage (%) (figure 4.2.3-1). Quel que soit le nombre de classes prises en considération, les valeurs représentées sur la diagonale descendante de la matrice correspondent aux instances correctement classifiées, tandis que les autres cellules correspondent aux confusions.

Prédit / Référence	Classe A	Classe B
Classe A	Vrais positifs (VP)	Faux négatifs (FN)
Classe B	Faux positifs (FP)	Vrais négatifs (VN)

Figure 4.2.3-1 : exemple de matrice de confusion à deux classes.

Les valeurs de précision et rappel sont obtenues grâce à ces valeurs de vrais et faux positifs, et vrais et faux négatifs.

La précision est la proportion d'instances qui font effectivement partie de cette classe dans laquelle ils ont été classés, soit $\frac{VP}{VP+FP}$. Le rappel est la proportion d'instances

d'une classe qui sont effectivement classées dans leur classe de référence, soit $\frac{VP}{VP+FN}$. La F-mesure, quant à elle, est une moyenne harmonique qui permet d'évaluer le rapport moyen entre la précision et le rappel.

Lorsque l'on fait une classification avec un plus de deux de classes, on a une valeur de précision, rappel et F-mesure pour chaque classe, qu'on peut ensuite combiner pour obtenir des valeurs correspondant à la tâche de classification dans sa globalité.

Nous avons utilisé la validation croisée, qui est une technique d'échantillonnage permettant d'estimer la fiabilité d'un modèle. Nous avons choisi une validation croisée à dix plis, c'est-à-dire que nous prenons dix échantillons comme ensemble de validation et les autres échantillons constitueront l'ensemble d'apprentissage. Nous calculons le score de performance, puis nous répétons l'opération en sélectionnant un autre échantillon de validation parmi les 9 échantillons qui n'ont pas encore été utilisés pour la validation du modèle.

4.2.3.1 ZeroR

L'algorithme ZeroR ne prenant aucune règle de classification, il classe, de ce fait, toutes les instances dans une seule et même classe, celle la plus représentée, ici la classe des 20-29 ans. Il faut alors dépasser le score de classification obtenu, ce score fait office de *baseline*.

Pour nos données, le score de classification est de 53 %, soit 411 instances bien classées sur 775. Ce qui montre que nos classes ne sont pas équilibrées en nombre, puisque sinon, le score à battre serait 17% (1 / nombre de classes du corpus, soit 1/6). Étant donné que nous avons à la fois des hommes et des femmes dans notre corpus, nous prenons en compte séparément les *baselines* de chaque sexe, soit 51 % pour les hommes et 54% pour les femmes.

Le score de classification du sexe avec ZeroR est de 56%, ce score constituera donc la *baseline* pour la classification du sexe.

4.2.3.2 JRIP

L'algorithme JRip (RIPPER) examine les classes en taille croissante et un ensemble initial de règles pour la classe est généré à l'aide du procédé d'erreur réduite incrémentielle de JRip (RIPPER) en traitant tous les exemples d'un jugement particulier dans les données comme une catégorie, et trouver un ensemble de règles qui couvrent tous les membres de cette catégorie. Ensuite, l'algorithme passe à la classe suivante et procède de la même manière et ainsi de suite jusqu'à ce que toutes les classes aient été traitées (Rajput and Prasad Aharwal 2011).

4.2.3.2.1 MFCC

Pour le jeu de données total, nous obtenons un score de classification de 76%, avec des valeurs moyennes de précision, rappel et F-mesure égales à 0.766, 0.764 et 0.764.

Pour les femmes, le score de classification passe à 83.5 %, soit une amélioration de 7% par rapport à la classification avec les deux sexes réunis (tableau 4.2.3-1).

Tableau 4.2-1 : Matrice de confusion des taux d'instances prédites pour une classe en fonction de la référence pour les MFCC, avec l'algorithme JRIP pour les femmes.

Prédit / Référence	15-19	20-29	30-39	40-49	50-59	60-69	N_total
15-19	60 %	38 %	0 %	2 %	0 %	0 %	45
20-29	4 %	87 %	0 %	3 %	4 %	1 %	236
30-39	0 %	27 %	73 %	0 %	0 %	0 %	15
40-49	0 %	9 %	3 %	74 %	15 %	0 %	34
50-59	0 %	7 %	0 %	1 %	92 %	0 %	90
60-69	0 %	27 %	0 %	0 %	0 %	73 %	15

Pour les hommes, nous obtenons un score de classification de 85%, soit une amélioration de 9% par rapport au score de classification des deux sexes réunis, et un meilleur score que les données des femmes.

4.2.3.2.2 Descripteurs phonétiques

La classification en classe d'âge avec JRIP nous donne un score de 60 %, avec des valeurs moyennes de précision, rappel et F-mesure égales à 0.582 et 0.603 et 0.565. En supprimant les valeurs d'écart-type du ZCR des voyelles /i/ et /y/ et du formant F3, nous obtenons un score de 62% avec des valeurs moyennes de précision, rappel et F-mesure égales à 0.607, 0.619 et 0.585.

Pour les femmes, le score de classification en fonction de la classe d'âge est de 61%, avec des valeurs moyennes de précision, rappel et F-mesure égales à 0.591, 0.614 et 0.590. Soit une amélioration de 1% en comparaison avec la classification sur toutes les données réunies. Ce score monte à 63 % lorsque l'on supprime la valeur d'écart-type du formant F2, avec des valeurs moyennes de précision, rappel et F-mesure égales à 0.611, 0.632 et 0.610. Nous pouvons de plus observer que la majorité des confusions sont des instances prédites comme appartenant à la classe 20-29 ans, ce qui peut être équivalent à une erreur d'un âge médian de quarante ans pour la classe d'âge 60-69 ans qui a 20% de d'instances prédites en classe 20-29 ans (tableau 4.2.3-2).

Référence \ Prédit	15-19	20-29	30-39	40-49	50-59	60-69	Nombre total d'enregistrements
15-19	38 %	60 %	0 %	0 %	2 %	0 %	45
20-29	3 %	84 %	0 %	0 %	11 %	1 %	236
30-39	0 %	13 %	80 %	0 %	7 %	0 %	15
40-49	3 %	47 %	0 %	29 %	21 %	0 %	34
50-59	4 %	52 %	1 %	9 %	32 %	1 %	90

60-69	0 %	20 %	13 %	0 %	7 %	60 %	15

Tableau 4.2-2 : Matrice de confusion des taux d'instances prédites pour une classe en fonction de la référence pour les descripteurs phonétiques, avec l'algorithme JRIP pour les femmes.

Pour les hommes nous obtenons alors un score de classification en fonction de la classe d'âge de 64 % soit 2% au-dessus de la classification tous sexes confondus avec les paramètres acoustiques supprimés, et si nous supprimons la valeur d'écart-type des consonnes fricatives /z/ et /ʒ/, nous obtenons un score de classification de 65.5%. Nous avons donc amélioré la classification de 5% par rapport à la classification avec tous les deux sexes réunis. En ce qui concerne la distance entre la classe de référence et la classe prédite, nous observons le même phénomène que pour les femmes, avec de forts pourcentages de confusion pour chaque classe en classe d'âge 20-29 ans (tableau 4.2.3-3).

Prédit / Référence	20-29	30-39	40-49	50-59	60-69	N_total
20-29	83 %	2 %	2 %	11 %	2 %	175
30-39	73 %	4 %	16 %	7 %	0 %	45
40-49	24 %	4 %	64 %	7 %	0 %	45
50-59	52 %	3 %	2 %	43 %	0 %	60
60-69	20 %	0 %	0 %	20 %	60 %	15

Tableau 4.2-3 : Matrice de confusion des taux d'instances prédites pour une classe en fonction de la référence pour les descripteurs phonétiques, avec l'algorithme JRIP pour les hommes.

4.2.3.2.3 Combinaison des descripteurs

Pour le jeu de données complet, nous obtenons un score de classification de 76.5%, en ne conservant que les paramètres phonétiques qui optimisaient la classification, soit sensiblement le même score que la classification pour les MFCC seuls.

Pour les femmes, le score de classification en fonction de la classe d'âge est de 84.5% avec des valeurs moyennes de précision, rappel et F-mesure égales à 0.845, 0.846 et 0.843. La classification pour les femmes a donc été améliorée de 1% en combinant les paramètres phonétiques et les MFCC.

Nous observons que les confusions ont tendance à concerner des classes d'âge adjacentes, sauf pour la classe 60-69 ans pour laquelle nous avons 20% de confusion qui concerne la classe 20-29 ans soit un âge médian de différence de quarante ans (tableau 4.2.3-4).

Référence \ Prédit	15-19	20-29	30-39	40-49	50-59	60-69	N_total
15-19	64 %	33 %	0 %	2 %	0 %	0 %	45
20-29	2 %	91 %	0 %	2 %	4 %	1 %	236
30-39	0 %	13 %	73 %	0 %	13 %	0 %	15
40-49	0 %	9 %	0 %	68 %	24 %	0 %	34
50-59	2 %	11 %	0 %	0 %	87 %	0 %	90
60-69	0 %	20 %	0 %	0 %	0 %	80 %	15

Tableau 4.2-4 : Matrice de confusion des taux d'instances prédites pour une classe en fonction de la référence pour tous les descripteurs, avec l'algorithme JRIP pour les femmes.

Pour les hommes, nous obtenons un score de classification de 84.5% en retirant l'écart-type du ZCR des consonnes /z/ et /ʒ/ comme pour les descripteurs phonétiques seuls. Ce résultat est légèrement inférieur au résultat de la classification par les MFCC seuls.

Contrairement aux confusions des femmes, les confusions pour les hommes semblent concerner majoritairement des classes plus éloignées, comme pour la classe 60-69 ans dont 27% des confusions sont prédites en classe 40-49 ans donc âge médian de différence

de vingt ans, ou la classe 50-59 ans dont 13 des confusions sont prédites en classe 20-29 ans soit un âge médian de différence de trente ans (tableau 4.2.3-5).

Prédit / Référence	20-29	30-39	40-49	50-59	60-69	N_total
20-29	91 %	2 %	1 %	6 %	1 %	175
30-39	36 %	64 %	0 %	0 %	0 %	45
40-49	9 %	2 %	87 %	0 %	2 %	45
50-59	13 %	3 %	0 %	82 %	2 %	60
60-69	7 %	0 %	27 %	0 %	67 %	15

Tableau 4.2-5 : Matrice de confusion des taux d'instances prédites pour une classe en fonction de la référence pour tous les descripteurs, avec l'algorithme JRIP pour les hommes.

4.2.3.2.4 Classification en fonction du sexe

Nous avons réalisé la classification en fonction du sexe avec JRIP, afin de savoir si les classifications avec descripteurs phonétiques et les MFCC pouvaient être porteur d'information sur le sexe du locuteur.

Avec les MFCC comme descripteurs, nous obtenons un score de classification du sexe de 85%, pour les descripteurs phonétiques, nous avons obtenu un score de classification de 95.5%, soit 10% de plus que la classification avec les MFCC et pour tous les descripteurs combinés, le score de classification du sexe est de 94%, un score inférieur à la classification du sexe avec les descripteurs phonétiques (figure 4.2.3-2).

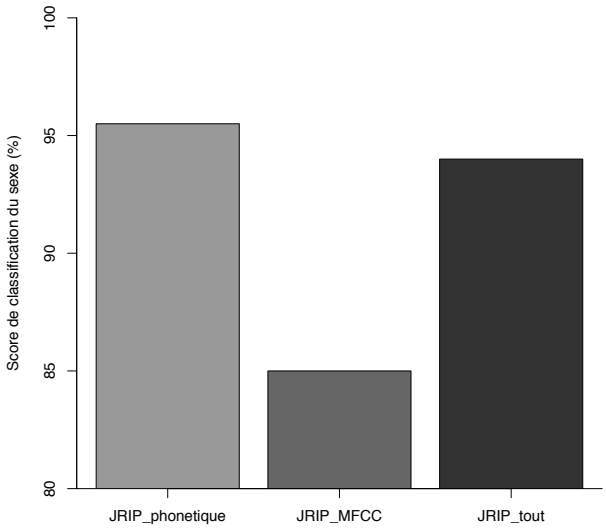


Figure 4.2.3-2 : Comparaison des moyennes de score de classification du sexe (%) fonction du type de descripteurs pour l'algorithme JRIP.

4.2.3.2.5 Bilan

La classification en fonction de la classe d'âge avec les MFCC comme descripteurs donne de meilleurs résultats que celle avec seulement les descripteurs phonétiques. Ces derniers semblent cependant être de meilleurs paramètres pour la classification en fonction du sexe que les MFCC.

La classification en fonction de l'âge sur le jeu de données des femmes d'un côté et celui des hommes de l'autre se trouve être un bon facteur d'amélioration de la classification, surtout lorsque les MFCC sont introduits dans la classification.

La combinaison des deux types de paramètres n'est pas, dans le cas de l'algorithme JRIP, un facteur stable de l'amélioration du score de classification, puis que l'ajout des paramètres phonétique à la classification des MFCC n'améliore le score que des locutrices (figure 4.2.3-3).

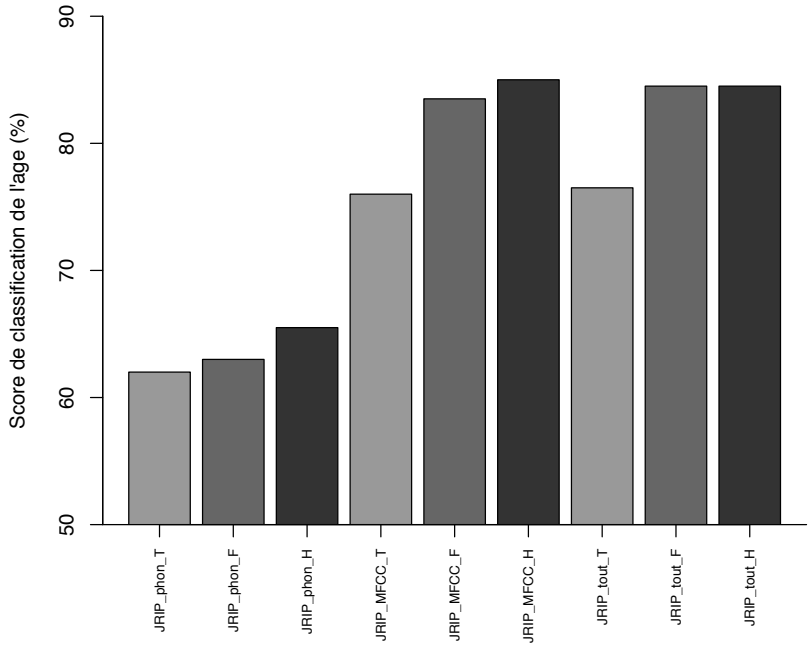


Figure 4.2.3-3 : Comparaison des moyennes de score de classification de l'âge (%) des hommes et des femmes en fonction du type de descripteurs pour l'algorithme JRIP (T= tous sexes confondus, F= femmes, H = hommes).

Les scores de classification de la classe d'âge ont tendance à être meilleurs pour les hommes-3que pour les femmes, mais il faut noter que le jeu de données des hommes comporte une classe d'âge de moins que le jeu de données des femmes, cela peut avoir une influence sur la classification. De plus, il faut rappeler que nous avons un effectif de locuteurs masculins qui est moindre que celui des femmes, ce qui réduit la variabilité intra-classe pour les hommes.

L'amélioration de la classification par les descripteurs phonétiques ne s'est pas faite par la suppression du même paramètre pour les femmes, pour lesquelles nous avons supprimé la valeur d'écart-type du formant F2 et pour les hommes, pour lesquels nous avons supprimé la valeur d'écart-type du ZCR des consonnes fricatives /z/ et /ʒ/ (tableau 4.2.3-6).

	JRIP			JRIP	
Descripteur	F	H	Descripteur	F	H
Moyenne de F0	x	x	Durée moyenne des fricatives	x	x
sdF0	x	x	Durée moyenne des voyelles	x	x
Moyenne de F1	x	x	Durée totale de l'énoncé	x	x
sdF1	x	x	ZCR moyen de / i/ et /y/	x	x
Moyenne de F2	x	x	sdZCR de / i / et /y/	x	x
sdF2		x	ZCR moyen de /z/ et /ʒ/	x	x
Moyenne de F3	x	x	sdZCR de /z/ et /ʒ/	x	
sdF3	x	x	ZCR moyen de /f/ et /s/	x	x
CGS moyen	x	x	sdZCR de /f/ et /s/	x	x
SdCGS	x	x	Nombre de descripteurs	18	18

Tableau 4.2-6 : Récapitulatif des paramètres phonétiques conservés ou non, pour l'algorithme JRIP en fonction du sexe.

4.2.3.3 J-48

La méthode J-48 est basée sur l'algorithme « C4.5 » développé par Ross Quinlan, mais dont les droits sont réservés. Il s'agit d'un algorithme établissant des arbres de décision qui tente de discriminer les instances selon leur classe, en fonction des attributs qui semblent meilleurs que les autres (Haccoun 2012).

L'objectif de ce type de méthode est de construire une fonction de classement représentable par un arbre qui est construit en partant de la racine et en allant vers les feuilles. On cherche à discriminer les exemples selon leur classe et en fonction d'attributs considérés comme les meilleurs parmi tous les autres au sens d'un critère donné. J48 est une implémentation open source de l'algorithme C4.5.

4.2.3.3.1 MFCC

Avec la classification J-48 et les MFCC comme descripteurs nous obtenons un taux d'instances correctement classifiée de 77%, avec une précision, un rappel et une F-mesure respectivement égaux à 0.774, 0.775 et 0.774.

Pour les femmes, nous obtenons un score de classification de 81 %avec des valeurs moyennes de précision, rappel et F-mesure égale à 0.817, 0.814, 0.814, soit un score meilleur que toutes les données mélangées de 4%.

Nous pouvons observer de forts taux de confusions, parfois pour des classes d'âge assez éloignées, comme pour la classe 60-69 ans dont 20% de confusion concernent la classe 20-29 ans, soit un âge médian de quarante ans (tableau 4.2.7-7).

Prédit / Référence	15-19	20-29	30-39	40-49	50-59	60-69	N_total
15-19	67 %	24 %	2 %	2 %	4 %	0 %	45
20-29	3 %	85 %	3 %	2 %	6 %	1 %	236
30-39	0 %	20 %	73 %	0 %	7 %	0 %	15
40-49	0 %	18 %	0 %	71 %	12 %	0 %	34
50-59	2 %	8 %	2 %	3 %	84 %	0 %	90
60-69	0 %	20 %	0 %	0 %	0 %	80 %	15

Tableau 4.2-7 : Matrice de confusion des taux d'instances prédites pour une classe en fonction de la référence pour les descripteurs MFCC, avec l'algorithme J-48 pour les femmes.

Pour les hommes, nous obtenons un score de classification de 85.9 %, avec des valeurs moyennes de précision, rappel et F-mesure égales à 0.860, 0.859 et 0.859, des résultats meilleurs que ceux des femmes de 4 % et meilleurs que ceux des deux sexes combinés de 8 %.

4.2.3.3.2 Descripteurs phonétiques

Tous sexes confondus, la classification avec les paramètres phonétiques donne un taux d'instances correctement classifiées de seulement 56% avec des valeurs de précision, rappel et F-mesure de seulement 0.559, 0.560, 0.559, ce qui signifie que nous sommes très proches du hasard obtenu avec ZeroR (53%).

Référence	15-19	20-29	30-39	40-49	50-59	60-69	N_total
15-19	51 %	38 %	0 %	0 %	11 %	0 %	45
20-29	4 %	69 %	5 %	7 %	15 %	0 %	411
30-39	0 %	43 %	32 %	7 %	15 %	3 %	60
40-49	0 %	32 %	6 %	47 %	14 %	1 %	79
50-59	3 %	41 %	5 %	11 %	38 %	3 %	150
60-69	3 %	17 %	13 %	7 %	10 %	50 %	30

Tableau 4.2-8 : Matrice de confusion des taux d'instances prédites pour une classe en fonction de la référence pour les descripteurs phonétiques, avec l'algorithme J-48.

Nous avons modulé les paramètres afin de voir si en en enlevant certains, les résultats pourraient être meilleurs. Effectivement, nous avons observé de meilleurs résultats en supprimant les valeurs d'écarts-types de la fréquence fondamentale et des formants F1, F2, et F3, les valeurs moyennes de ZCR, sauf celles des voyelles i et y, ainsi que les valeurs d'écarts-types, les valeurs moyennes de durée des voyelles. Nous obtenons alors un pourcentage d'instances bien classées de 58.326%, avec des valeurs de précision, rappel et F-mesures égales à 0.571, 0.583 et 0.575.

Pour les femmes, nous obtenons un score de classification de l'âge de 59.54%, avec des mesures de précision, rappel et F-mesure de 0.585, 0.595 et 0.589, valeurs que nous avons tenté d'améliorer en supprimant certains descripteurs. Ainsi, en supprimant les valeurs moyennes et d'écart-type du ZCR des voyelles /i/ et /y/, les valeurs d'écart-type du

CGS et des formants F1, F2 et F3, ce score remonte à 63.22% avec des valeurs moyennes de précision, rappel et F-mesure égales à 0.631, 0.632 et 0.629.

Nous pouvons observer, dans le tableau 4.2.3-9, que les confusions ne concernent pas toujours des classes adjacentes, par exemple, la classe 60-69 ans à un taux de confusion en classe 15-19 ans de 13%, soit un âge médian de différence égal à 47.5 ans.

Prédit / Référence	15-19	20-29	30-39	40-49	50-59	60-69	N_total
15-19	42 %	49 %	0 %	2 %	4 %	2 %	45
20-29	8 %	74 %	1 %	3 %	14 %	0 %	236
30-39	0 %	27 %	60 %	0 %	0 %	13 %	15
40-49	3 %	35 %	0 %	26 %	35 %	0 %	34
50-59	3 %	36 %	0 %	12 %	47 %	2 %	90
60-69	13 %	0 %	20 %	13 %	13 %	40 %	15

Tableau 4.2-9 : Matrice de confusion des taux d'instances prédites pour une classe en fonction de la référence pour les descripteurs phonétiques, avec l'algorithme J-48, pour les femmes.

Pour les hommes, nous obtenons un score de 58.82% avec des valeurs moyennes de précision, rappel et F-mesure égales à 0.592, 0.588 et 0.588. En supprimant les valeurs moyenne de durée des voyelles ainsi que les valeurs moyennes et d'écarts-types du CGS et enfin les valeurs d'écart-type de la F0 et du formant F2, nous obtenons finalement un score de 60.6%. Ce qui reste inférieur aux résultats des femmes, mais meilleur que si les données des deux sexes étaient confondues.

Du tableau 4.2.3-10, nous pouvons observer les mêmes observations que pour les femmes, en ce qui concerne la distance en classe d'âge entre les classe de référence et les confusions, en notant que les locuteurs ont tendance à être classés plus jeunes que leur âge réel, la classe 30-39 ans par exemple a un taux d'instances prédites dans la classe d'âge 20-29 ans plus fort que celui dans sa classe réelle, si nous prenons les valeurs d'âge

médian des classes de référence et prédite, nous avons 44% d'énoncés de locuteurs qui sont classés dans une classe dont l'âge médian est plus petit de dix ans. 24% productions de certains locuteurs de la classe 40-49 ans sont prédits dans la classe 20-29 ans soit une différence médiane de vingt ans. Enfin, 32% des productions des locuteurs de la classe 50-50 ans et 20% des productions des locuteurs de la classe 60-69 ans sont classés dans une classe avec un âge médian de différence de trente ans, respectivement les classes 20-29 ans et 30-39 ans.

Référence \ Prédit	20-29	30-39	40-49	50-59	60-69	N_total
20-29	73 %	9 %	8 %	9 %	1 %	175
30-39	44 %	36 %	9 %	9 %	2 %	45
40-49	24 %	7 %	64 %	4 %	0 %	45
50-59	32 %	13 %	5 %	43 %	7 %	60
60-69	13 %	20 %	7 %	13 %	47 %	15

Tableau 4.2-10 : Matrice de confusion des taux d'instances prédites pour une classe en fonction de la référence pour les descripteurs phonétiques, avec l'algorithme J-48, pour les hommes.

4.2.3.3.3 Combinaison des descripteurs

En combinant les paramètres acoustiques et les MFCC, le taux d'instances correctement classifiées passent à 78,19 % avec des valeurs de précision, rappel et F-mesure respectivement égales à 0.777, 0.782 et 0.779, soit des valeurs légèrement supérieures à celles de la classification avec seulement les MFCC.

Pour les femmes, nous obtenons un score de classification de 83 %, avec des valeurs moyennes de précision, rappel et F-mesure de 0.835, 0.830 et 0.831.

Nous observons une tendance des confusions des locutrices des classes 50-59 ans et 60-69 ans à concerner des classes d'âge assez éloignées (tableau4.2.3-11).

Référence \ Prédit	15-19	20-29	30-39	40-49	50-59	60-69	N_total
15-19	73 %	20 %	4 %	2 %	0 %	0 %	45
20-29	3 %	87 %	2 %	2 %	6 %	1 %	236
30-39	0 %	20 %	73 %	0 %	7 %	0 %	15
40-49	0 %	18 %	0 %	71 %	12 %	0 %	34
50-59	0 %	9 %	2 %	4 %	84 %	0 %	90
60-69	0 %	13 %	7 %	0 %	0 %	80 %	15

Tableau 4.2-11 : Matrice de confusion des taux d'instances prédites pour une classe en fonction de la référence pour tous les descripteurs, avec l'algorithme J-48 pour les femmes.

Pour les hommes, le score de classification avec les MFCC et les descripteurs phonétiques conservés pour améliorer les résultats, nous obtenons un taux d'instances correctement classifiées de 86.7 %, et des valeurs moyennes de précision rappel et F-mesure égales à 0.867, 0.868 et 0.867. Il s'agit d'une amélioration de 1% par rapport aux MFCC seuls et une amélioration de 3 % par rapport aux femmes.

Référence \ Prédit	20-29	30-39	40-49	50-59	60-69	N_total
20-29	91 %	2 %	1 %	6 %	0 %	175
30-39	16 %	80 %	0 %	4 %	0 %	45
40-49	2 %	4 %	89 %	0 %	4 %	45
50-59	17 %	7 %	0 %	77 %	0 %	60
60-69	0 %	0 %	13 %	0 %	87 %	15

Figure 4.2.3-4 : Matrice de confusion des taux d'instances prédites pour une classe en fonction de la référence pour tous les descripteurs, avec l'algorithme J-48, pour les hommes.

4.2.3.3.4 Classification en fonction du sexe

Les descripteurs phonétiques ayant été très performant pour la classification du sexe pour l'algorithme JRip, nous nous sommes intéressés aux résultats de cette même classification avec l'algorithme J48.

Le score de classification du sexe avec les MFCC est de 86%, pour les descripteurs phonétiques ce score monte 96.2%, soit 746 instances bien classées sur 775, et pour tous les descripteurs combinés, le score de classification du sexe redescend à 94.6%

Pour l'algorithme J48, il semblerait que les critères phonétiques soient plus efficaces pour la classification en fonction du sexe que les MFCC, qui font baisser le score de classification du sexe si on les combine avec les descripteurs phonétiques (figure 4.2.3-5).

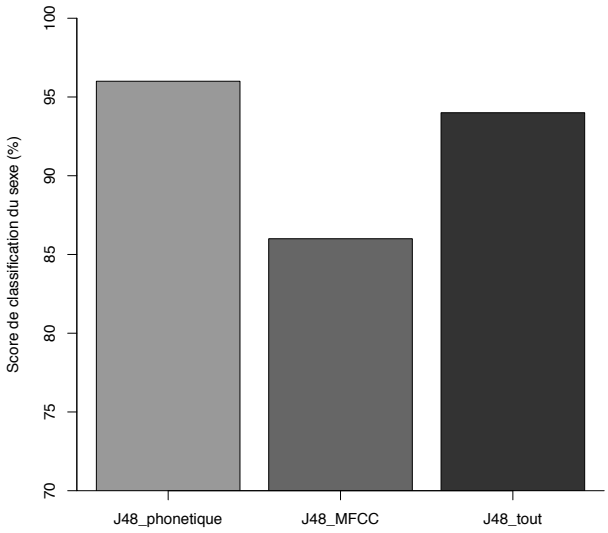


Figure 4.2.3-5 : comparaison des scores de classification par le sexe en fonction des descripteurs pour l'algorithme J48.

4.2.3.3.5 Bilan

Séparer les données des hommes et des femmes nous a de nouveau permis d'améliorer les scores de classification.

La classification est meilleure lorsque l'on utilise les MFCC comme descripteurs que lorsque nous utilisons les descripteurs phonétiques seuls, cependant l'extraction de ces derniers n'est pas vaine pour autant car un fois combinés aux MFCC, ils permettent d'améliorer que quelques pourcents les résultats de la classification (figure 4.2.3-6).

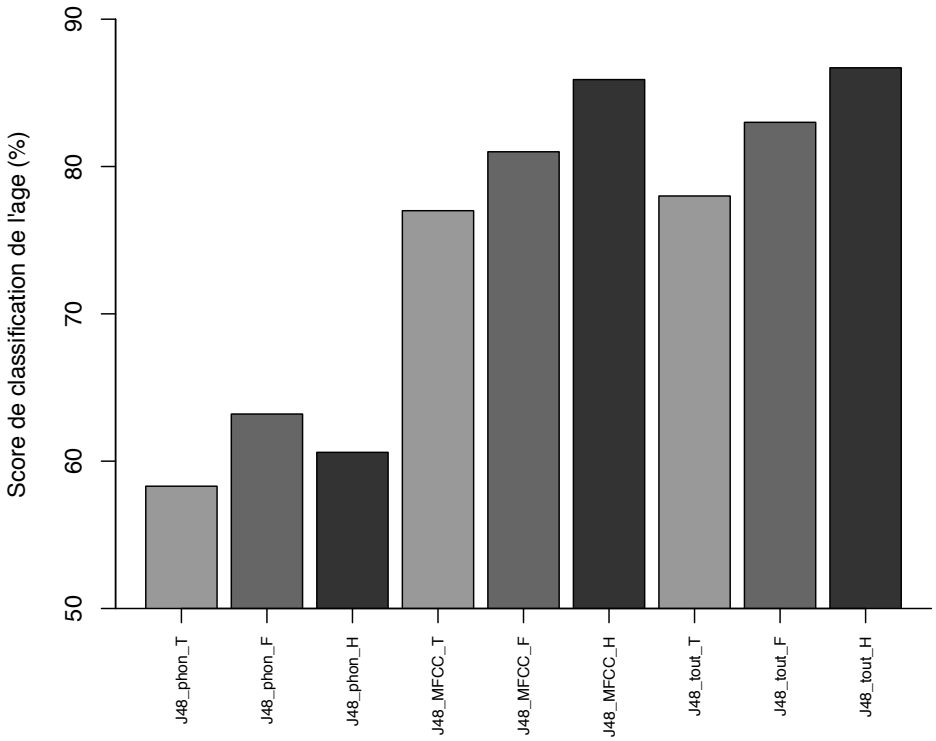


Figure 4.2.3-6 : Comparaison des moyennes de score de classification (%) des hommes et des femmes en fonction du type de descripteurs pour l'algorithme J48 (T= tous sexes confondus, F= femmes, H = hommes).

Nous pouvons de plus nous rendre compte que les descripteurs phonétiques qui ont été retiré pour les hommes et pour les femmes ne sont pas totalement identiques. En effet, seuls huit descripteurs sur les dix-neuf extraits permettent d'améliorer les scores de classification de l'âge pour les deux sexes (tableau 4.2.3-12).

	J48			J48	
Descripteur	F	H	Descripteur	F	H
Moyenne de F0	x	x	Durée moyenne des fricatives	x	x
sdF0	x		Durée moyenne des voyelles	x	
Moyenne de F1	x	x	Durée totale de l'énoncé	x	x
sdF1		x	ZCR moyen de /i/ et /y/		x
Moyenne de F2	x	x	sdZCR de /i/ et /y/		x
sdF2			ZCR moyen de /z/ et /ʒ/	x	x

Moyenne de F3	x	x	sdZCR de /z/ et /ʒ/	x	
sdF3		x	ZCR moyen de /f/ et /s/	x	x
CGS moyen	x		sdZCR de /f/ et /s/	x	x
SdCGS		x	Nombre de descripteurs	13	14

Tableau 4.2-12 : Récapitulatif des paramètres phonétiques conservés ou non, pour l'algorithme J48 en fonction du sexe

4.2.3.4 SMO (Sequential Minimal Optimization)

Le SMO est un algorithme développé par John C. Platt pour entrainer un classifieur à vecteur de support (SVM). Cet algorithme va découper un problème de programmation quadratique en une série de problèmes de programmation quadratique plus petits (C. Platt 1999)

4.2.3.4.1 Descripteurs MFCC

Tous sexes confondus, nous obtenons un résultat de 87% avec des valeurs de précision, rappel et F-mesure de 0.877, 0.874 et 0.868.

Pour les femmes, nous avons donc un score de classification de la classe d'âge de 92 % soit 403 instances bien classées sur 435, avec des valeurs moyennes de précision, rappel et F-mesure égales à 0.926, 0.926 et 0.925. Nous avons donc amélioré la classification de 5%.

Prédit / Référence	15-19	20-29	30-39	40-49	50-59	60-69	N_total
15-19	76 %	24 %	0 %	0 %	0 %	0 %	45
20-29	1 %	95 %	0 %	1 %	3 %	0 %	236
30-39	0 %	0 %	100 %	0 %	0 %	0 %	15
40-49	0 %	6 %	0 %	91 %	3 %	0 %	34
50-59	0 %	7 %	0 %	0 %	93 %	0 %	90
60-69	0 %	0 %	0 %	0 %	0 %	100 %	15

Tableau 4.2-13 : Matrice de confusion des taux d'instances prédites pour une classe en fonction de la référence pour les MFCC, avec l'algorithme SMO chez les femmes.

Pour les hommes, nous obtenons un résultat de classification de la classe d'âge de 98.2% MFCC avec des valeurs moyennes de précision, rappel et F-mesure égales à 0.982. Ce qui équivaut à une amélioration de la classification de 11%.

Référence \ Prédit	20-29	30-39	40-49	50-59	60-69	N_total
20-29	98 %	1 %	0 %	1 %	0 %	175
30-39	4 %	96 %	0 %	0 %	0 %	45
40-49	0 %	0 %	100 %	0 %	0 %	45
50-59	2 %	0 %	0 %	98 %	0 %	60
60-69	0 %	0 %	0 %	0 %	100 %	15

Tableau 4.2-14 : Matrice de confusion des taux d'instances prédites pour une classe en fonction de la référence pour les MFCC, avec l'algorithme SMO chez les femmes.

Nous avons donc une amélioration moyenne de la classification par la classe d'âge de 8%.

4.2.3.4.2 Descripteurs phonétiques

Pour les descripteurs phonétiques, nous obtenons un pourcentage d'instances bien classées de 58.97%, avec des valeurs de précision, rappel et F-mesure qui ne sont, cependant, pas exploitables car aucune instance appartenant à la classe 30-39 ans n'a été classées dans cette classe (VP=0), et aucune instance n'y a été classée par erreur (FP=0). En supprimant les valeurs de ZCR sauf pour /i/ et /y/, ainsi que les valeurs d'écarts-types des valeurs moyennes de ZCR, nous obtenons un résultat de classification de 59.742%, les valeurs de précision, rappel et F-mesure ne sont toujours pas exploitables pour la classe 30-39 ans.

Pour les données des locutrices, avec les descripteurs phonétiques, nous obtenons un score de classification de 63.45%, c'est un résultat qui est 4% au-dessus de la même classification sur les données avec hommes et femmes confondus et en ayant supprimé quelques paramètres. Nous n'avons, cependant, pas observé d'amélioration de score de classification en supprimant certains paramètres phonétiques. Nous avons de plus des classes pour lesquelles les valeurs de précision rappel et F-mesure sont nulles ou très faibles, par exemple la classe 15-19 ans (rappel = 0.178, F-mesure = 0.291) et la classe 30-39 ans (précision non déterminée, rappel nul).

Référence \ Prédit	15-19	20-29	40-49	50-59	60-69	N_total
15-19	18 %	82 %	0 %	0 %	0 %	45
20-29	1 %	95 %	0 %	4 %	0 %	236
30-39	0 %	73 %	0 %	27 %	0 %	15
40-49	0 %	50 %	3 %	47 %	0 %	34
50-59	0 %	57 %	0 %	41 %	2 %	90
60-69	0 %	33 %	0 %	33 %	33 %	15

Tableau 4.2-15 : Matrice de confusion des taux d'instances prédites pour une classe en fonction de la référence pour les descripteurs phonétiques, avec l'algorithme SMO chez les femmes.

Pour les hommes, nous obtenons un résultat de classification par la classe d'âge de 60.9%, soit un résultat meilleur que ce que nous avions pour tous sexes confondus. Mais nous pouvons améliorer cette classification en supprimant les valeurs d'écart-types des ZCR et du formant F3, ainsi que la durée moyenne des fricatives. Nous obtenons ainsi un score de classification de 63%.

Référence \ Prédit	20-29	30-39	40-49	50-59	60-69	N_total
20-29	93 %	0 %	3 %	3 %	1 %	175

30-39	76 %	7 %	11 %	7 %	0 %	45
40-49	42 %	0 %	53 %	4 %	0 %	45
50-59	63 %	0 %	0 %	37 %	0 %	60
60-69	40 %	7 %	0 %	33 %	20 %	15

Tableau 4.2-16 : Matrice de confusion des taux d'instances prédites pour une classe en fonction de la référence pour les descripteurs phonétiques, avec l'algorithme SMO chez les hommes.

4.2.3.4.3 Combinaison des descripteurs

En combinant les descripteurs MFCC et les descripteurs phonétiques (sauf ceux qui donnaient de moins bon résultats), nous obtenons un score de classification de 87.613% contre 87.23 % soit 679 instances bien classées sur 775, avec des valeurs de précision, rappel et F-mesure moyennes égales à 0.880, 0.876 et 0.871.

Prédit Référence	15-19	20-29	30-39	40-49	50-59	60-69	N_total
15-19	78 %	22 %	0 %	0 %	0 %	0 %	45
20-29	0 %	95 %	1 %	0 %	3 %	0 %	411
30-39	0 %	53 %	43 %	0 %	3 %	0 %	60
40-49	0 %	8 %	0 %	87 %	5 %	0 %	79
50-59	0 %	15 %	0 %	0 %	85 %	0 %	150
60-69	0 %	3 %	0 %	0 %	0 %	97 %	30

Figure 4.2.3-7 : Matrice de confusion des taux d'instances prédites pour une classe en fonction de la référence pour tous les descripteurs, avec l'algorithme SMO, tous sexes confondus.

Il est intéressant de noter que les classes 15-19 ans, 40-49 ans et 60-69 ans ont une précision supérieure à 0.958 et que la plupart des erreurs de classification concerne une classe d'âge adjacente, à l'exception de la classe 50-59 ans pour laquelle les instances mal classées sont reportée en classe 20-29 ans.

Pour les locutrices, nous obtenons un score de classification de l'âge de 94.48%, soit 7% au-dessus du score des deux sexes combinés et 2% au-dessus des MFCC seuls.

Pour les hommes, nous obtenons un score de classification en fonction de la classe d'âge de 98.2%, soit exactement le même pourcentage pour les MFCC seuls, cependant, nous observons quelques améliorations au niveau des valeurs de précision, rappel et F-mesure au détail de la classe, avec l'ajout des paramètres phonétiques, nous avons maintenant trois classes d'âge avec des valeurs de précision, rappel et F-mesure égales à 1, contre deux sans les paramètres phonétiques. Il s'agit des classes 40-49 ans, 50-59 ans, 60-69 ans.

4.2.3.4.4 Classification du sexe

Comme pour les deux algorithmes précédents, nous avons essayé de classer les instances en fonction du sexe du locuteur.

Pour les MFCC, nous obtenons un score de classification du sexe de 94.6 %, avec les descripteurs phonétiques est de 96.77%, soit 750 instances bien classées sur 775 et pour tous les descripteurs combinés nous obtenons un score de classification de 98.9 % d'instances bien classée, soit 767 instances sur 775.

Avec l'algorithme SMO, les critères phonétiques semblent plus efficaces pour la classification en fonction du sexe, mais, contrairement aux algorithmes précédents, la combinaison des deux types de descripteurs permet d'améliorer de manière notable la classification (figure 4.2.3-8).

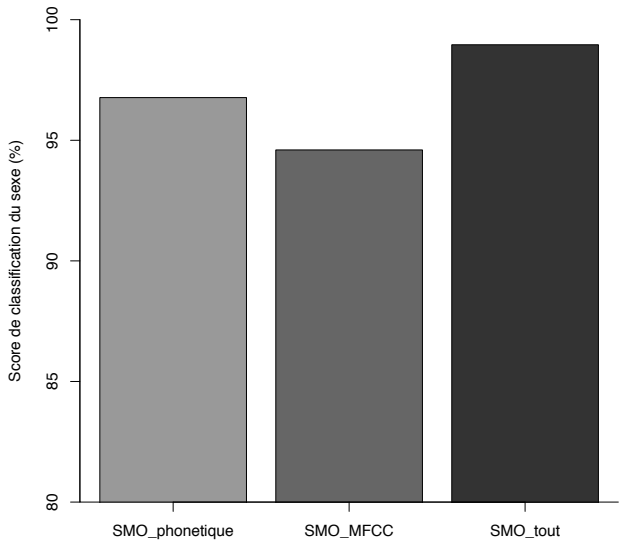


Figure 4.2.3-8 : comparaison des scores de classification par le sexe en fonction des descripteurs pour l'algorithme SMO.

4.2.3.4.5 Bilan

Quels que soient les descripteurs, nous remarquons encore que la classe avec la meilleure valeur de rappel est la classe 20-29 ans (0.866, tous descripteurs confondus) et la classe avec la moins bonne valeur de rappel est la classe 30-39 ans (0.467, tous descripteurs confondus), cela s'explique du fait que la première classe est la plus représentée du corpus, et la deuxième est la moins représentée.

Séparer les données des hommes et des femmes nous a permis d'améliorer les scores de classification (figure 4.2.3-9). La classification est meilleure lorsque l'on utilise les MFCC comme descripteurs que lorsque nous utilisons les descripteurs phonétiques seuls, cependant l'extraction de ces derniers n'est pas vaine pour autant car un fois combinés aux MFCC, ils permettent d'améliorer de plusieurs pourcents les résultats de la classification.

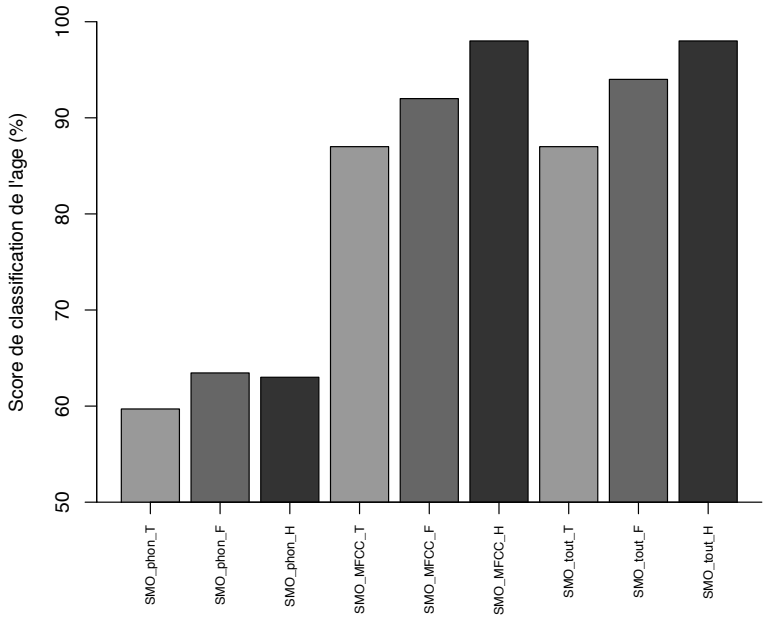


Figure 4.2.3-9 : Comparaison des moyennes de score de classification (%) des hommes et des femmes en fonction du type de descripteurs pour l'algorithme SMO (T= tous sexes confondus, F= femmes, H = hommes).

Nous pouvons de plus nous rendre compte que les descripteurs phonétiques qui ont été retiré pour les hommes et pour les femmes ne sont pas totalement identiques, et le nombre de descripteurs pour maximiser la classification est passé à 14 au lieu de 19 pour les deux sexes.

	SMO			SMO	
Descripteur	F	H	Descripteur	F	H
Moyenne de F0	x	x	**Durée moyenne des fricatives**	x	
sdF0	x	x	**Durée moyenne des voyelles**	x	x
Moyenne de F1	x	x	**Durée totale de l'énoncé**	x	x
sdF1	x	x	**ZCR moyen de /i/ et /y/**		x
Moyenne de F2	x	x	**sdZCR de /i/ et /y/**		
sdF2	x	x	**ZCR moyen de /z/ et /ʒ/**		x
Moyenne de F3	x	x	**sdZCR de /z/ et /ʒ/**		
sdF3	x		**ZCR moyen de /f/ et /s/**		x

CGS moyen	x	x	sdZCR de /f/ et /s/		
SdCGS	x	x	Nombre de descripteurs	14	14

Tableau 4.2-17: Récapitulatif des descripteurs phonétiques conservés, en fonction du sexe pour le modèle SMO, avec le nombre total de descripteurs conservé

5 DISCUSSIONS ET CONCLUSIONS

5.1 CONCLUSIONS

Nous vous avons présenté seulement la classification des données enregistrées en ligne, car les données enregistrées en présentiel étaient moins bien réparties en termes de classes d'âge et de sexes. De plus, plusieurs enregistrements qui avaient été réalisés sont bruités par une climatisation qui s'entend et s'observe au spectrogramme, appliquer une classification sur ces données auraient peut-être classifié le « bruité » / « non-bruité » plutôt que la classe d'âge, notamment pour les coefficients cepstraux MFCC, pour lesquels nous ne connaissons pas la provenance des paramètres.

Nous avons pu classifier nos données selon trois algorithmes qui ont des fonctionnements différents et nous obtenons les résultats de classification suivants (tableau 5.1-1).

La classification en fonction des descripteurs phonétique seuls est moins bonne que la classification, des MFCC seuls, mais la combinaison des descripteurs phonétiques et des MFCC peut permettre d'améliorer jusqu'à 2.5% la classification de l'âge. Il faudra noter que c'est pour la classification de l'âge pour les locutrices qui est la plus bénéficiaire d'une combinaison des MFCC et des descripteurs phonétiques.

De plus, pour nos données, le modèle le plus efficace est le SMO, pour lequel le score de classification avec les MFCC avant même de les avoir combinés avec les

descripteurs phonétiques est supérieur à 90%, score que les deux autres modèles n'ont pu atteindre même une fois maximisés (tableau 5.1-1).

	JRIP		J48		SMO	
	F	H	F	H	F	H
MFCC	83	85	81	86	92	98
Phonétique	63	65,5	63	60	63,5	63
Combinaison	84,5	84,5	83	87	94,5	98
Apport	1,5	-0,5	2	1	2,5	0
ZeroR (*baseline)*	F = 54, H = 51					

Tableau 5.1-1 : Bilan des scores de classification de la classe d'âge en fonction du sexe et de l'algorithme de classification.

Nous pouvons alors conclure que les descripteurs phonétiques ne sont pas de meilleurs classifieurs que les MFCC comme nous pouvions nous en douter mais ils nous permettent tout de même de savoir quelles caractéristiques sont des indices de vieillissement de la voix pour nos données. En effet, du tableau 5.2.3, nous apprenons que les descripteurs qui ont le plus de poids dans la classification de la classe d'âge du locuteur sont les valeurs moyennes de F0, F1, F2 et F3 ainsi que la durée totale de la production qui nous donne des informations sur le débit. Ce sont des résultats auxquels nous pouvions nous attendre, puisque ce sont les principaux changements liés à l'âge relatés dans la littérature.

Descripteurs	JRIP		J48		SMO		occurence
	F	H	F	H	F	H	
Moyenne de F0	x	x	x	x	x	x	100
Écart-type de F0	x	x	x		x	x	83
Moyenne de F1	x	x	x	x	x	x	100
Écart-type de F1	x	x		x	x	x	83
Moyenne de F2	x	x	x	x	x	x	100
Écart-type de F2		x			x	x	50
Moyenne de F3	x	x	x	x	x	x	100
Écart-type de F3	x	x		x	x		67
CGS moyen	x	x	x		x	x	83
Écart-type de CGS	x	x			x	x	67
Durée moyenne des fricatives	x	x	x	x	x		83
Durée moyenne des voyelles	x	x	x		x	x	83
Durée totale de l'énoncé	x	x	x	x	x	x	100
ZCR moyen de /i/ et /y/	x	x		x	x	x	83
Ecart-type du ZCR de /i/ et /y/	x	x		x			50
ZCR moyen de /z/ et /ʒ/	x	x	x	x		x	83
Ecart-type du ZCR de /z/ et /ʒ/	x		x	x			50
ZCR moyen de /f/ et /s/	x	x	x	x		x	83
Ecart-type du ZCR de /f/ et /s/	x	x	x	x			67
Total descripteurs (nombre)	18	18	13	14	14	14	

Tableau 5.1-2 : Récapitulatif des descripteurs utilisés en fonction du modèle et du sexe, avec le nombre total de descripteurs utilisés par modèle (Total descripteurs) et la proportion de modèle qui conserve le descripteur concerné)

5.2 DISCUSSIONS

5.2.1 Nos résultats

De précédentes lectures, nous avions appris que la fréquence fondamentale chez les femmes avait tendance à diminuer avec l'âge, alors que pour les hommes, la fréquence fondamentale aurait plutôt tendance à augmenter. Nos résultats sont donc assez surprenants puisque nous observons une F0 montante pour les femmes de plus de soixante ans, quelle que soit la condition d'enregistrement, et une F0 légèrement montante pour les hommes enregistrés en ligne.

Les valeurs de ZCR que nous avions obtenus pour les sons voisés, qui indiquaient des sons voisés plus périodiques pour les locuteurs plus âgés, sont un résultat auquel nous ne nous attendions pas. Etant donné que quelque soit le phone, les valeurs de ZCR ont le même contour d'évolution, nous pouvons, par exemple, nous demander si ce résultat ne reflèterait pas une meilleure qualité d'enregistrement chez les sujets plus âgés, qu'on ne peut pas directement évaluer avec le SNR en raison des post-traitements appliqués aux enregistrements en ligne.

Il aurait été intéressant de voir si les valeurs moyennes en Hz de F1, F2 et F3 pour les locuteurs les plus âgés correspondent toujours à des /ɛ/ et des /a/ ou si nous avons tendance à basculer sur d'autres catégories de voyelles, par exemple en comparant aux valeurs listées dans (Georgeton et al. 2012).

5.2.2 Améliorations possibles

Il aurait, de plus, été intéressant de pousser la variation intralocuteur à la différence de réalisation des différentes phrases, plus dans le détail pour une même condition. Nous pourrions nous demander si l'ordre dans lequel les deux conditions sont enregistrées a une influence, si le locuteur commence par les enregistrements en conditions contrôlées, il y

aurait peut-être moins de chances qu'il nous soumette des enregistrements incomplets ou invalides, car il aura eu un contact avec l'expérimentateur qui aura pu lui expliquer les consignes d'une différente manière et le locuteur pourra peut-être être un meilleur juge de la qualité de ses propres enregistrements en ligne par la suite.

Étant donné que nous avons montré que certains descripteurs phonétiques pouvaient avoir un poids plus fort que les autres pour la classification de l'âge, nous pourrions tester la classification avec des sous-ensembles de paramètres phonétiques : uniquement ceux relatifs à F0, uniquement ceux relatifs aux formants, uniquement les durées, etc., pour lesquels la proportion de modèles qui les conservait était égale à 100%. Nous aurions aimé classifier les données du panel de locuteurs qui avait enregistré à la fois en présentiel et en ligne, afin de voir si les enregistrements d'un même locuteur quelle que soit la condition d'enregistrement étaient prédits dans la même classe.

5.2.3 Recommandations pour utilisations futures d'enregistrement en ligne

Comme nous avons pu le voir en section 3.1 – Recueil de données, nous avons mis en place deux processus de recueil de données et nous observons que dans nos deux sous-corpus, certaines classes peuvent être sous-représentées voire non représentées, qu'il y a des déséquilibres en termes de répartition des sexes. Les classes d'âges et sexes peuvent être rééquilibrées mais cela aurait demandé un temps de recueil plus large, or, nous n'avons pas réussi à mettre en place des stratégies de déploiement de la plateforme d'enregistrement assez conséquentes. Les données dont nous disposons ont donc principalement été recueillies auprès de proches et de proches de proches.

L'enregistrement de données en ligne rallonge le temps de recueil des enregistrements et réduit le contrôle que l'expérimentateur exerce sur le locuteur enregistré. En effet, l'expérimentateur n'a la main sur les données seulement après l'envoi

des données sur le serveur et n'a pas de contrôle sur le matériel d'enregistrement ou les conditions d'enregistrement, il peut seulement faire un tri ou filtrage a posteriori, ce qui entraine la perte de certaines données incomplètes ou inexploitables. Il est donc très important de détailler au maximum ce qui est attendu de l'utilisateur car il ne pourra pas interagir avec l'expérimentateur par un autre biais que les consignes qui lui sont laissées sur le site.

Il faut faire preuve de souplesse et s'adapter à tout type d'utilisateur, en prenant compte que la maîtrise de l'outil informatique n'est pas excellente pour tous les utilisateurs et certains auront besoin de consignes très développées.

Liste des tableaux

Liste des figures

Bibliographie

- Bachu, R. G., B. Adapa, S. Kopparthi, and B. D. Barkana. 1978. “Separation of Voiced and Unvoiced Using Zero Crossing Rate and Energy of the Speech Signal.” *Electrical Engineering Department School of Engineering, University of Bridgeport* 128.
- Boersma, P. 1993. “Accurate Short-Term Analysis of the Fundamental Frequency and the Harmonics-to-Noise Ratio of a Sampled Sound.” Pp. 97–100 in *Proceedings of the institute of phonetic sciences*. Vol. 17, *1193*.
- C. Platt, John. 1999. “Using Analytic QP and Sparseness to Speed Training of Support Vector Machines.”
- Colletta, Jean-Marc, Catherine Pellenq, and Isabelle Rousset. 2008. “Evolution du débit de parole chez l’enfant francophone dans des tâches narrative et conversationnelle. 27èmes Journées d’Etudes sur la Parole, Association Francophone de la Communication Parlée.”
- Cornut, Guy. 2009. “La voix parlée, Chapitre II.” Pp. 43–58 in *La voix*.
- D. Reichel, Uwe. 2012. “PermA and Balloon: Tools for String Alignment and Text Processing.”
- Davis, S. B. and P. Mermelstein. 1980. “Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences.”
- Deshpande, Hrishikesh, Rohit Singh, and Unjung Nam. 2001. “Classification of Music Signals in the Visual Domain.” Limerick, Ireland.
- Georgeton, Laurianne, Nikola Paillereau, Simon Landron, Jiayin Gao, and Takeki Kamiyama. 2012. “Analyse Formantique Des Voyelles Orales Du Français En Contexte Isolé : À La Recherche d’une Référence Pour Les Apprenants de FLE.” Pp. 145–52 in, *ffhalshs-00977591f*. Grenoble, France.
- Haccoun, Adrien. 2012. “Comparaison de Méthodes de Classifications.”
- Haton, Jean-Paul, Christophe Cerisara, Dominique Fohr, Yves Laprie, and Kamel Smaïli. 2006. “Reconnaissance Automatique de la Parole Du signal à son interprétation.” P. 392 in. Paris: UniverSciences.

- Jollois, François-Xavier. 2003. “Contribution de La Classification Automatique à La Fouille de Données : Ordinateur et Société.” *Université Paul Verlaine - Metz*.
- Kahn, Juliette. 2011. “Parole de locuteur : performance et confiance en identification biométrique vocale.” Informatique, Université d’Avignon et des Pays de Vaucluse, Avignon.
- Lévêque, N., M. Laganaro, C. Fougeron, V. Delvaux, M. Pernon, S. Borel, and S. Catalano. 2016. “MonPaGe: Un Protocole Informatisé d’évaluation de La Parole Pathologique En Langue Française.”
- Martin, Philippe. 2008. *Phonétique Acoustique : Introduction à l’analyse Acoustique de La Parole*. Paris: A. Colin.
- Meigner, Sylvain and Mickael Rouvier. 2012. “Nouvelle Approche Pour Le Regroupement Des Locuteurs Dans Des Émissions Radiophoniques et Télévisuelles.” Pp. 97–104 in *ATALA*. Vol. 1. Grenoble, France: AFCP.
- Meunier, Christine. 2007. “Phonétique acoustique.” Pp. 164–73 in *Les dysarthries*.
- Moritz, Niko, Kamil Adiloglu, Jörn Anemüller, Stefan Goetze, and Birger Kollmeier. 2016. “Multi-Channel Speech Enhancement and Amplitude Modulation Analysis for Noise Robust Automatic Speech Recognition.” *Computer Speech & Language* 46:558,573.
- Nam, Unjung. 2001. “Special Area Exam Part II.”
- Phuoc, Nguyen, Le Trung, Xu Huang, and Sharma Dharmendra. 2010. “Fuzzy Support Vector Machines for Age and Gender Classification.”
- Rabiner, L., S. Levinson, A. Rosenberg, and J. Wilpon. 1979. “Speaker-Independent Recognition of Isolated Words Using Clustering Techniques.” in Vol. asp-27, *no.4*.
- Rajput, Anil and Ramesh Prasad Aharwal. 2011. “J48 and JRIP Rules for E-Governance Data.” P. 201 in Vol. 5.
- Santini, Vincent. 2016. “Impact perceptuel d’une mise à zéro des segments plosifs de parole.” UNIVERSITÉ DE SHERBROOKE, Faculté de génie Département de génie électrique et de génie informatique, Sherbrooke, Canada.
- Schiel, Florian. 1999. “Automatic Phonetic Transcription of Non-Prompted Speech.” Pp. 607–10 in. San Francisco.
- Schötz, Susanne. 2007. “Analysis and Synthesis of Speaker Age.”

- Schötz, Suzanne. 2006. “Perception, Analysis and Synthesis of Speaker Age.” *Linguistics and Phonetics.*
- Shepstone, Sven Ewan, Zheng-Hua Tan, and Søren Holdt Jensen. 2013. “Audio-Based Age and Gender Identification to Enhance the Recommendation of TV Content.” *IEEE Transactions on Consumer Electronics* 59:721–29.
- Simon, Anne-Catherine. 2007. “Guide méthodologique - Transcription outillée, prosodies.”
- Tellier, Isabelle. n.d. “Introduction à La Fouille de Textes.”
- Torre III, Peter and Jessica A. Barlow. 2009. “Age-Related Changes in Acoustic Characteristics of Adult Speech.” *San Diego State University, United States*.
- Vaissière, Jacqueline. 2015. “Le signal de parole et la phonétique acoustique, Chaptitre V.” P. 61 in *La phonétique*. puf.

Annexe 1 : Préparation des données en ligne à l'utilisation de WebMaus Général

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Fri May  3 12:16:27 2019
@author: audreygombault
Script qui permet de generer un fichier par qui contient la transcription SAMPA
de l'énoncé produit afin de pouvoir appliquer MAUS général
Le fichier produit à exactement le même nom que le fichier wav
"""
import os
print(os.getcwd())
folder_path = "donnees_enligne/fichiers/"
for path, dirs, files in os.walk(folder_path):
    for filenames in files:
        if ".DS_Store" not in filenames and "phrase" in filenames and filenames
.endswith(".wav"):
            ph_num = filenames.split("_")[-1].split(".")[0][-2:]
            # on coupe en _ puis on garde que le dernier bout qu'on sépare en .
et
            # dont on ne garde que les deux derniers caractères de la première
partie
            inputfile = open("phrases/par/phrase_{}.par".format(ph_num))
            newfile = open("donnees_enligne/fichiers/{}.par".format(filenames.s
plit(".")[0]),"w")
            for line in inputfile:
                newfile.write(line)
    newfile.close()
```

Annexe 2 : Appariement des fichiers wav et de leur Textgrid

```
#!/usr/bin/env python3
# -*- coding: utf-8 -*-
"""
Created on Fri May  3 11:54:13 2019
@author: audreygombault
Script qui crée un fichier texte comprenant les noms des fichiers wav et le fic
hier textgrid associé afin de pouvoir utiliser le script Praat d'extraction de
paramètres acoustiques de Mr Nicolas Audibert
"""
import os
folder_path = "../TEXTGRIDS_ENLIGNE/"
file = open("fichiersapparies_enligne1.txt", "w")
file.write("wav\tTextGrid\n")
for path, dirs, files in os.walk(folder_path):
    for filenames in files: # pour tous les noms de fichiers
        if ".DS_Store" not in filenames:
            wavfile = filenames.split('.')[0]+".wav"
            file.write(wavfile)
            file.write('\t')
            file.write(filenames)
            file.write("\n")
    file.close()
```

Annexe 3 : Traitement des données de comparaison

```
library(readr)
library(tidyr)
library(dplyr)
library(stringr)
setwd("/Users/audreygombault/MASTER/M2STAGE/DONNEES/R/")
enligne_fichier_parametres_acoustiques_niveau_phone = "R_ENLIGNE/ANALYSES/an
alyse_v13_ENLIGNE2_InfosLoc.txt"
presentiel_fichier_parametres_acoustiques_niveau_phone = "R_MIC/ANALYSES/ana
lyse_v13_tousphones_micro_InfosLoc.txt"
parametres_acoustiques_enligne = read_tsv(enligne_fichier_parametres_acousti
ques_niveau_phone, na = c("", "NA", "None"))
parametres_acoustiques_presentiel = read_tsv(presentiel_fichier_parametres_a
coustiques_niveau_phone, na = c("", "NA", "None"))

fichier_durees_enligne = "totalduration.csv"
dureeavecpauses_enligne = read_tsv(fichier_durees_enligne)
fichier_durees_presentiel = "totalduration_mic.csv"
dureeavecpauses_presentiel = read_tsv(fichier_durees_presentiel)
noms_textgrid_recodes = str_replace(dureeavecpauses_presentiel$textgrid_file
, "_reduced", "")

dureeavecpauses_presentiel$textgrid_file = sapply(
  noms_textgrid_recodes,
  function(x) as.character(x[1])
)

categoriephone = read_tsv("categories_phones.txt")
parametres_acoustiques_enligne = parametres_acoustiques_enligne %>% left_joi
n(categoriephone,by=c("label"="segmentSAMPA"))
parametres_acoustiques_presentiel = parametres_acoustiques_presentiel %>% le
ft_join(categoriephone,by=c("label"="segmentSAMPA"))
parametres_acoustiques_enligne = parametres_acoustiques_enligne %>% left_joi
n(dureeavecpauses_enligne,by=c("textgrid_file"="textgrid_file"))
parametres_acoustiques_presentiel = parametres_acoustiques_presentiel %>% le
ft_join(dureeavecpauses_presentiel,by=c("textgrid_file"="textgrid_file"))
parametres_acoustiques_enligne$condition <- "enligne"
parametres_acoustiques_presentiel$condition <- "presentiel"

codecontroles = read_tsv("../paired_result_code2.txt")
parametres_acoustiques_enligne = parametres_acoustiques_enligne %>% left_joi
n(codecontroles)
parametres_acoustiques_enligne$age_locuteur[parametres_acoustiques_enligne$c
ode_controle=="HER"]<-58
donneesenligne_codecontrole = parametres_acoustiques_enligne %>%
  filter(code_controle != "NA")
parametres_acoustiques_presentiel = parametres_acoustiques_presentiel %>% le
ft_join(codecontroles)
donneespresentiel_codecontrole = parametres_acoustiques_presentiel %>%
  filter(code_controle != "NA")

names(donneesenligne_codecontrole) <- names(donneespresentiel_codecontrole)
donnees_codecontrole = rbind(donneesenligne_codecontrole, donneespresentiel_
codecontrole)
```

Annexe 4 : Extraction des MFCC (Notebook Python)

```
import numpy, path, os, csv, praatio
from python_speech_features import mfcc
from python_speech_features import logfbank
import scipy.io.wavfile as wav
import python_speech_features
import praatio, os, path
from praatio import tgio

path_to_textgrid = "TEXTGRIDS_ENLIGNE/"

folderpath = "donnees_enligne/"
for path, dirs, files in os.walk(folderpath):
    for filenames in files :
        if filenames.endswith(".wav"):
            (rate,sig) = wav.read(folderpath+filenames)
            for chemins, directions, fichiers in os.walk(path_to_textgrid):
                for nomfichiers in fichiers:
                    if ".DS_Store" not in nomfichiers:
                        if nomfichiers.split('.')[0] == filenames.split('.')[0]:
                            tg = tgio.openTextgrid(path_to_textgrid+nomfichiers)
                            entryList = tg.tierDict["KAN-MAU"].entryList
                            # Get all intervals
                            startpoint = entryList[0][0]
                            startpoint1 = (startpoint*rate)
                            endpoint = entryList[-1][1]
                            endpoint1 = (endpoint*rate)
                            mfcc = python_speech_features.base.mfcc(sig[int(startpoint1)
:int(endpoint1)], samplerate=rate, numcep=19, nfilt=38, nfft = 1200)# winlen = 0.025, wi
nstep = 0.010 = 10ms
                            numpy.savetxt("MFCC/{}.txt".format(filenames.split(".")[0]),
mfcc, delimiter="\t", header="c0\tc1\tc2\tc3\tc4\tc5\tc6\tc7\tc8\tc9\tc10\tc11\tc12\tc13
\tc14\tc15\tc16\tc17\tc18", comments='')
```

Nous vous rappelons que plusieurs annexes sont disponibles en ligne à l'adresse https://github.com/AudreyGombault/Annexes.