UNIVERSITE PARIS III-SORBONNE NOUVELLE
Institut de Linguistique et Phonétique Générales et Appliquées

MEMOIRE DE MASTER 2 Traitement Automatique des Langues
parcours Recherche & Développement

présenté par Gabriele CHIGNOLI

Reconnaissance Automatique du Locuteur, mécanisme humain et tâche informatique : application de méthodes statistiques.

Sous la direction de
Cédric GENDROT

ANNEE UNIVERSITAIRE 2017-2018

Table des matières

Liste des figures

Liste des tableaux

Liste des abréviations

ACP Principal Component Analysis | Analyse en Composantes Principales

API Application Programming Interface

CART Classification and Regression Tree | Arbre de Classification et Regression

DNN Deep Neural Network | Réseaux Neurones Profonds

dB Décibel

EER Equal Error Rate | Taux d'Erreur Moyen

ERM Empirical Risk Minimisation | Minimisation Empirique du Risque

f_0 fréquence fondamentale

fMLLR feature-space Maximum Likelihood Linear Regression | espace de paramètre selon Régression Linéaire à Ressemblance Majeure

GMM-UBM Gaussian Mixture Model - Universal Background Model | Modèle à Mélange de Gaussiennes - Modèle du Monde

HASR Human Assisted Speaker Recognition | Reconnaissance Assisté du Locuteur

Hz Hertz

HMM Hidden Markov Model | Modèle de Markov Caché

HNR Harmonics to Noise Ratio | Rapport Harmoniques sur bruit

IAL Identification Automatique du Locuteur

LFCC Linear Frequency Cepstral Coefficients | Coefficients Cepstraux à Fréquence Linéaire

LIA Laboratoire d'Informatique d'Avignon

LPP Laboratoire de Phonétique et Phonologie

LRE Language Recognition Evaluation | Évaluation de la Reconnaissance de la Langue

MFCC Mel Frequency Cepstral Coefficients | Coefficients Cepstraux à Fréquence Mel

NIST National Institut of Standard and Technology

PC*n* Principal Component *n* | Composante Principale *n*

PLP Perceptual Linear Predictive | Prédiction Linéaire Perceptuelle

PVI Pairwise Variability Index | Indice de Variabilité Apparié

RAL Reconnaissance Automatique du Locuteur

SL Suivi du Locuteur

SM Spectral Moments | Moments Spectraux

SMO Sequential Minimal Optimisation | Optimisation Séquentielle Minimale

SNR Signal to Noise Ratio | Rapport Signal Bruit

SRE Speaker Recognition Evaluation | Évaluation de la Reconnaissance du Locuteur

SRM Structural Risk Minimisation | Minimisation Structurale du Risque

SVM Support Vector Machine | Machines à Vecteurs de Support ou Séparateurs à Vaste Marge

TTS Text To Speech

VAD Voice Activity Detector | Détection d'Activité Vocale

VAL Vérification Automatique du Locuteur

η^2 Eta Squared | Êta carré

Résumé

Ce mémoire s'inscrit dans le domaine de la Reconnaissance Automatique du Locuteur, une tâche informatique liée au traitement de la parole dans laquelle un système fournit une réponse, binaire ou selon un score de confiance, sur l'identité d'un locuteur dont il analyse un extrait sonore donné en entrée et appelé test. Nous avons choisi pour cette analyse de comparer les Coefficients Cepstraux à Fréquence Mel (MFCC), paramètres habituellement utilisés en Reconnaissance pour caractériser le locuteur, et un ensemble de mesures acoustiques : des indices spectraux et reliés au bruit dans le signal ; la durée et le rythme de l'élocution ; des mesures de fréquence fondamentale. L'analyse effectuée se base sur l'évaluation de la variabilité expliquée par ces indices au niveau phonémique et à l'aide de méthodes, telles que l'Analyse en Composantes Principales et les arbres de classification, pour tester la robustesse effective des indices face à la variabilité des données. Nous utilisons un corpus de voix journalistique et un deuxième composé d'enregistrements en milieu contrôlé.

Abstract

This study aims to do a comparative analysis, in the Automatic Speaker Recognition domain, between a set of Mel Frequency Cepstral Coefficients, the state of the art features used to describe a speaker in ASR, and some acoustic measurement such as : spectral and noise indication ; the speech duration and rhythm ; pitch values. A Speaker Recognition system takes one or two voice recording as input, named the test, depending on the answer it will create. A binary one if the speaker identity has to be extracted from two inputs, a trust threshold will determine the output in the other case. Our tests will come from controlled and uncontrolled environments in order to understand how the features react in different speaking situations. The features strength will be evaluated thank to statistical and classification methods and we will compare our results, when it will be possible, to similar works.

1 Introduction

Le traitement du langage parlé par ordinateur constitue une discipline en évolution continue, de nombreuses nouvelles techniques ont été introduites au fil des années que cela concerne la synthèse, la reconnaissance ou en générale l'analyse de productions langagières. L'objet principal de l'étude qui va suivre sera porté sur la Reconnaissance Automatique du Locuteur (RAL). Les travaux déjà effectués dans le domaine de la reconnaissance du locuteur sont basés majoritairement sur des coefficients cepstraux à fréquence Mel (MFCC) dont nous analyserons les avantages et les limites. Ces paramètres ne peuvent pas être reliés à des caractéristiques articulatoires, représentant une altération du signal sonore ils restent difficiles à interpréter de façon globale. Nous allons ici appliquer des mesures acoustiques phonétiques, plutôt liées à l'articulatoire, afin de tester leur efficacité en les comparant aux MFCC.

La voix est un flux d'air généré par les poumons, modulé par la cavité buccale et/ou nasale pour donner lieu à des sons plus ou moins intelligibles partageant des propriétés phonétiques. Notre voix possède des caractéristiques physiologiques liées à la façon dont notre corps est fait, auxquelles s'ajoutent des caractéristiques comportementales comme la vitesse d'élocution, les accents, les émotions, que nous pouvons plutôt relier à des indices paralinguistiques. Cet ensemble complexe forme la signature vocale. Dans la présente analyse, un autre pas menant à l'identification des éléments pertinents dans la variabilité inter locuteur de la signature vocale sera fait afin également de pousser encore plus loin les réflexions dans le domaine de la reconnaissance du locuteur.

La reconnaissance de la voix d'un locuteur essaye d'avoir un rôle central dans les systèmes biométriques ainsi que dans notre vie de tous les jours. Pour que ce soit possible deux éléments sont fondamentaux : une trace de test, c'est-à-dire un échantillon de la voix dont on veut obtenir ou identifier l'identité, extraite selon des paramètres ; une réponse obtenue en comparant le test avec des modèles d'une base déjà récoltée. Une donnée biométrique possède trois facteurs principaux (Chadha et al., 2011). L'**universalité**, la voix dans ce sens représente un élément très performant, les individus capables de produire des sons le font de la même manière et à travers les mêmes organes. De plus elle est simple à récolter : un locuteur qui souhaite utiliser sa voix à des fins de sécurité sera à priori très

collaboratif ; et le processus de stockage des données n'est pas invasive pour l'utilisateur. La **permanence** à travers le temps est aussi importante pour assurer l'efficacité face au locuteur. La **distinction** entre différents sujets, c'est sur ce dernier facteur que la recherche centre ses avancées. La reconnaissance du locuteur trouve sa place autant dans les applications biométriques que dans le milieu juridique, où elle a été fortement délimitée dans le passé, du moins en ce qui concerne la France (GCP, 1990 ; GFCP, 1999 ; AFCP, 2002), car selon les experts du domaine les systèmes actuels n'offrent pas encore une fiabilité assez élevée, nous reviendront sur ces questions au cours de l'analyse.

La première partie de ce mémoire consistera à faire le point sur l'état de l'art de la Reconnaissance du Locuteur et ses applications aujourd'hui. La deuxième partie sera dédiée à la description des corpus utilisés, un des principaux points de discussion dans le domaine de la RAL étant donné le peu de variabilité des bases de données soumises aux systèmes. Les outils de travail seront exposés dans la troisième partie. Enfin, dans les quatrième et cinquième parties les mesures utilisées ainsi que les résultats obtenus seront présentés afin de comprendre quels éléments sont des bons points d'analyse pour la suite des travaux.

2 État de l'art

"Every person uses speech forms in a unique way" (Bloomfield, 1933)

Cette phrase du linguiste Bloomfield dans son ouvrage Language de 1933 peut bien résumer l'idée à partir de laquelle cette étude a été conçue. Si toute personne produit de façon unique une forme de langage alors nous pouvons retrouver lors de l'analyse du discours à la fois des éléments qui marquent la variabilité et d'autres qui resteront constants. Pour les définir ce chapitre se basera sur la littérature en linguistique (Blumstein et Stevens, 1981 ; Pierrehumbert, 2003 ; Vaissière, 2006) afin de présenter une base de connaissances pour l'étude qui suivra, dans un second temps le domaine de la reconnaissance du locuteur ainsi que les tâches qui lui sont reliées seront abordées, pour continuer avec les approches et les méthodes d'analyse les plus répandues, in fine les perspectives posées surtout par les séries d'évaluation NIST.

2.1 Invariance et variabilité

La voix est le résultat du flux d'air généré par les poumons, modulé par la cavité buccale et/ou nasale et qui à partir des plis vocaux produit un son ensuite modulé par tout ce qui se présente et peut physiologiquement agir lors de la phonation (dents, lèvres, langues, palais, cavité nasale, etc.) en fonction des propriétés phonologiques que l'on souhaite reproduire ou autrement dit ce que l'on souhaite dire.

Comprendre le langage humain n'est pas un travail facile beaucoup de facteurs et de propriétés entrent en jeu. La compréhension du langage parlé cependant commence très tôt chez les humain, (Colombo et Bundy, 1983) ont observé comment les nouveaux nés peuvent dès 2 mois distinguer des sons de parole et des "bruits non linguistiques" et entre 5 et 6 mois ils commencent déjà à faire la distinction entre les mots de fonction et de contenu (Shi et Werker, 2001 ; 2003). (DeCasper et Fifer, 1980) ont aussi démontré la reconnaissance d'une voix familière par des enfants de deux à trois mois et ainsi leur capacité à la préférer à d'autres voix inconnues (Floccia et al., 2000). Dans le cas de voix féminines les enfants se retournent plutôt vers celle de leur mère, ils peuvent également reconnaitre leur langue maternelle uniquement à partir de la prosodie (Mehler et al., 1988)

ou du timbre (Floccia et al., 1997). Ces capacités de reconnaissance et d'analyse d'un signal sonore autant développés chez les humains font l'objet d'études linguistiques et représentent depuis l'arrivée des ordinateurs le modèle à atteindre. Les machines n'agissent pas de la même façon que le cerveau humain. Dans le prochain paragraphe le fonctionnement des systèmes automatiques de reconnaissance sera présenté, pour le moment nous allons exposer quelques notions propres à l'analyse et à la perception humaines du langage.

Pour permettre la reconnaissance du locuteur, du côté humain ou ordinateur, seront nécessaires : l'invariance permettant l'étude d'un même trait ou segment qui soit cohérent dans toutes ses réalisations ; la variabilité pour ce facteur qui sera représentée dans notre cas par les productions de locuteurs différents. À l'intérieur du discours l'analyse des facteurs expliquant la première seront utilisés et des mesures permettant mieux d'expliquer la deuxième seront identifiées.

Premièrement nous allons définir le concept d'invariance comme l'idée de base de cohérence dans les productions orales. Ce concept se rapproche de la notion de perception catégorielle (Vaissière, 2006 ; Bidelman et al., 2013), lors de l'écoute d'un extrait de parole notre cerveau segmente automatiquement le discours en entrée et catégorise les segments par similarité avec d'autres segments qu'il connaît ou à défaut il crée de nouvelles catégories à partir de ce fait linguistique. Ceci peut se produire lorsque le segment, ou même la voix dans un contexte de reconnaissance de locuteur, est inconnu (Bidelman et Lee, 2015).

L'idée d'invariance dans la littérature a largement été discutée. (Eskenazi, 1984 ; Hombert et Puech, 1984 ; Serrurier et al., 2017) décrivent les trois raisons principales pour lesquelles l'invariance se trouve être un concept difficile. (1) Alors que nous parlons, nos articulateurs sont fortement dépendants des configurations précédentes et successives au segment cible, par conséquent les traits de cette unité seront influencés aussi par le contexte. (2) Il peut y avoir à tout moment lors de la phonation l'influence de traits qui ne représentent pas un contraste de la langue. La nasalité en anglais est évoquée à titre d'exemple (Blumstein, et Stevens 1981 ; Vaissière, 1986) : un locuteur lorsqu'il produit une voyelle suivie d'une consonne nasale est libre d'anticiper la nasalisation de la consonne en nasalisant la voyelle précédente ou non sans créer de problèmes de perception. (3) À des locuteurs différents correspondent différents appareils vocaux, en longueur et forme, il peut donc facilement y avoir une variation dans les fréquences naturelles du conduit vocal d'un locuteur à un autre

pour un même son. Cependant l'invariance fait partie des éléments rendant possible le processus de perception grâce à la construction d'un ensemble de modèles avec des propriétés définies qui se distribuent habituellement sur une fenêtre de fréquences de quelque dizaine de millisecondes.

La recherche de propriétés acoustiques invariantes correspondant à un trait phonétique, requiert, du point de vue du signal, une détermination non pas en termes de fréquence absolue d'un pic spectral particulier mais plutôt en termes de forme du spectre (Blumstein et Stevens, 1981) à des moments particuliers, pour l'homme il s'agit de retrouver la distance minimale entre le modèle construit et la propriété qui intervient dans le discours.

De l'autre côté nous avons la variabilité, qui peut être aussi bien intra qu'inter locuteur, cette notion permet d'établir la distance, en termes de différence entre ce qui est dit et ce qui est connu ou modélisé dans le discours prononcé (Greenberg, 2003). Ces deux notions sont en quelque sorte complémentaires et seront reprises tout au long de cette étude. Elles peuvent également être définies selon les concepts d'alignement et de distance, des concepts importants aussi bien en informatique qu'en phonétique (Wieling et al., 2012). Pouvoir calculer la distance de prononciation entre différentes réalisations d'une même unité, objet de l'invariance, qui peut être un mot, un phonème ou un simple trait phonologique, joue un rôle important par exemple pour définir des variantes dialectales. L'invariance prend la forme plutôt de variabilité lorsque des facteurs tels que le contexte, la prosodie, la sphère sociale, le style surviennent (Keating, 1997). À travers les observations sur les stratégies de nasalisation des différents locuteurs, (Vaissiere, 1995) rapporte une mineur variabilité intra locuteur dans la hauteur du voile du palais en posant l'accent sur les nasales comme source d'invariance et en même temps de variabilité inter locuteur. Ceci a aussi été observé par (Kahn, 2011) dans des tests de perception lors de la campagne NIST-HASR. Une échelle de discrimination a pu être établie pour les segments présentant une majeure invariance et un effet plus important sur la variabilité interlocuteur : voyelles nasales, consonnes nasales, voyelles orales (avec les segments [-fermé ; -postérieur] qui présentent moins d'influence), fricatives, plosives et approximantes. Une deuxième échelle est établie à partir des résultats obtenus en analysant les MFCC (Mel Frequency Cepstral Coefficients | Coefficients Cepstraux à Fréquence Mel) pour les mêmes segments : diphtongues, fricatives, nasales (voyelles et consonnes avec le même score), voyelles orales, plosives.

Les facteurs peuvent donc influencer l'invariance et la variabilité mais aussi représenter des champs d'étude importants pour établir les points forts de l'analyse de l'oral, la prosodie dans ce sens en fait pleinement partie, ce facteur s'est déjà montré très important dans la facilitation de la reconnaissance d'une langue (Ohala et Gilbert, 1979 ; Fougeron et Keating, 1995). Dans (Borràs-Comes et al., 2014) les auteurs considèrent les variations de fréquence fondamentale ou encore (Solé, 2003) met l'accent sur les traits phonologiques comme élément pour définir la variation et l'invariance dans les productions orales. Encore d'autres paramètres paralinguistiques comme l'orientation sexuelle du sujet font l'objet d'analyse dans (Pierrehumbert et al., 2004), cette étude se base sur le principe qu'à travers l'analyse du discours un grand nombre d'informations sur le locuteur peuvent être retrouvées.

2.2 RAL et voisins

La littérature du traitement automatique de la parole (Calliope, 1989) distingue traditionnellement deux grandes branches dans la reconnaissance automatique du locuteur (RAL) : l'identification (IAL) et la vérification (VAL). Les deux tâches peuvent être appliquées de deux manières différentes (Boujelbene et al., 2009). À travers des systèmes dépendants, où la trace prononcée pour le test est connue ou même proposée par l'évaluateur, ou à travers des systèmes indépendants du texte dans lesquels le test est totalement libre, ce qui donne lieu à des approches souvent assez différentes.

(Kahn, 2011) parle de segmentation en locuteurs, (Jousse, 2011) du suivi de locuteur (SL), ces deux tâches se situent dans une étape plus avancée de RAL qui relie IAL, VAL et l'effet du cocktail party. Ce mélange représente encore une limite importante pour les systèmes de reconnaissance voire, dans certaines situations, pour l'humain (Remez et al., 1997), cependant des solutions commencent à apparaitre (Wang et Sun, 2017). Comme tout système de traitement automatique de la langue, ceux de RAL comprennent une entrée, une série de traitements et une sortie. Pour la tâche d'Identification l'entrée est constitué par une simple trace, un extrait sonore du locuteur à identifier, ou par un couple de traces lorsque la tâche consiste à comprendre si les deux enregistrements appartiennent au même locuteur. Pour la Vérification une identité aussi sera donnée en entrée. L'IAL fournira une identité en sortie, obtenue en comparant le test aux modèles de locuteurs déjà présents dans sa base.

Un système de VAL donnera en sortie une réponse binaire, sur laquelle un score de confiance est appliqué pour définir en quelle mesure la réponse est acceptable, nous parlons de comparaison cible ou imposteur s'il s'agit du même locuteur pour les deux traces ou non et de faux rejet ou fausse acceptation dans les deux cas d'erreurs.

2.3 Approches et techniques

Nous allons maintenant explorer l'ensemble des traitements nécessaire pour passer de l'entrée à la sortie dans un système de RAL en exposant les techniques principales utilisées dans les travaux du domaine et en pointant les éléments qui seront importants pour la suite de l'étude. Les étapes de traitement comprennent tout d'abord une transformation du signal ou des signaux sonores de test à travers des paramètres déterminés par des techniques différentes et ensuite la comparaison du résultat de cette transformation avec les modèles de locuteurs déjà présents dans la mémoire du système, ensemble d'entrainement, et qui auront été obtenus, eux aussi, selon une approche différente ou similaire que le test.

2.3.1 Extraction des paramètres

(Davis et Mermelstein, 1980) décrivent comment la représentation et paramétrisation du signal acoustique peuvent être effectuées à travers plusieurs méthodes répondant principalement à deux approches : celle par spectre de Fourier et celle par spectre à prédiction linéaire. La première a donné lieu aux deux représentations qui recueillent le succès majeur en reconnaissance : les LFCC (Linear Frequency Cepstral Coefficients | Coefficients Cepstraux à Fréquence Linéaire) et les MFCC, présentant une échelle linéaire en dessous des 1000 Hertz (Hz) et logarithmique au-dessus, les mesures statistiques sur ces paramètres se trouvent être plus efficaces (Kamruzzaman et al., 2010).

Les coefficients cepstraux se sont imposés dans le domaine de la RAL (Brummer et Swart, 2014 ; Ahmad et al., 2015) mais l'approche linéaire a permis le calcul de paramètres à travers la technique dite analyse PLP (Perceptual Linear Predictive | Prédiction Linéaire Perceptuelle) (Hermansky, 1990). Un de ses résultats les plus populaires est le fMLLR (feature-space Maximum Likelihood Linear Regression | espace de paramètre selon

Régression Linéaire à Ressemblance Majeure) qui est retrouvé aussi dans des travaux récents de reconnaissance du locuteur (Khosravani et al., 2016 ; Zhang, 2018). Toutefois sa réussite reste inférieure en dehors de l'approche en sous-bandes.

L'extraction est effectuée habituellement sur des trames extraites chaque 30ms à partir de la trace principale sur laquelle un premier filtre peut être appliqué à travers un algorithme de VAD (Voice Activity Detection | Détection d'Activité Vocale) afin de garder les parties du signal avec le plus d'énergie à l'intérieur du vecteur acoustique. Des techniques modernes de fenêtrage (Sahidullah et Saha, 2013) ou de filtrage du signal en plusieurs sous-bandes ont montré aussi leur apport de résultats intéressants. Ceci a notamment été vu chez (Safavi et al., 2012) et (Mahola et al., 2007) où le système à sous-bandes proposé avec paramétrage non linéaire arrive à obtenir des résultats plus significatifs qu'un système de reconnaissance traditionnel en live test.

2.3.2 Modélisation du locuteur

Une fois le vecteur extrait à partir de la trace de test le système procède à la création d'un modèle du locuteur qui sera plus tard comparé à d'autres modèles déjà présents et produits de la même façon lors de l'entraînement.

La technique la plus répandue en RAL pour la modélisation est représentée par le i-vectors, basé sur le modèle linéaire de Gaussiennes (voir paragraphe suivant, GMM-UBM). Le locuteur est caractérisé selon une approche descriptive ce qui facilite les mesures statistiques sur ces vecteurs qui obtiennent de meilleurs résultats sur des échantillons plus courts (Li et al., 2016). Un autre moyen de représentation des vecteurs acoustiques est celui nommé d-vectors. Cette représentation est moins répandue et plus récente puisqu'elle se relie aux techniques des réseaux neurones (Richardson et al., 2015). Dans cette approche le locuteur est représenté en gardant uniquement les variances les plus pertinentes, la phase d'entrainement devient déterminante pour que le système apprenne quels sont les traits discriminants.

Encore d'autres techniques sont proposées dans les travaux de (Cai et al., 2018), principalement à travers des réseaux neurones, ils représentent peut-être les futurs développements dans le domaine. Une tendance à créer des systèmes hybrides est attestée

(Trabelsi et Ayed, 2013) pour maximiser la quantité d'informations contenues dans les vecteurs, ce qui demande un travail plus important par le système mais qui pourra garantir une discrimination plus fine pour certains paramètres, notamment les indices paralinguistiques (Schötz, 2002) comme l'accent ou la prosodie qui pourraient être impliqués dans la modélisation d'un locuteur provenant d'une certaine région ou pays.

2.3.3 Prise de décision

La prise de décision est la phase finale d'un système de reconnaissance du locuteur et ici les techniques pour le calcul du score de similarité entre la trace de test et les modèles d'entraînement ne manquent pas non plus. L'approche par réseaux neurones (Richardson et al., 2015) est la plus récente dans le domaine. Elle demande un travail supplémentaire puisque la modélisation (choix des traits discriminants) et la décision (calcul de la différence maximale entre les d-vecteurs) sont effectuées de manière différente, elle est surtout retrouvée avec des systèmes hybrides.

Les trois autres approches des systèmes de RAL sont le GMM -UBM (Reynolds, 1995 ; Reynolds et al., 2000) (Gaussian Mixture Model - Universal Background Model | Modèle à Mélange de Gaussiennes - Modèle du Monde), le SVM (Support Vector Machine | Machines à Vecteurs de Support ou Séparateurs à Vaste Marge) et le HMM (Hidden Markov Model | Modèle de Markov Caché). Le GMM-UBM occupe une position dominante, son fonctionnement est relié à la modélisation que les i-vectors effectuent. Pendant la phase d'entrainement la description en Gaussiennes de la distribution des paramètres acoustiques du locuteur favorise la prise de décision qui sera effectuée selon la majeure ressemblance du test avec les modèles.

En ce qui concerne le SVM la prise de décision est effectuée en séparant linéairement les vecteurs à travers un hyperplan pour après calculer la distance la plus courte entre le vecteur de test et ceux du modèle. Cette approche est une implémentation de la SRM (Structural Risk Minimisation | Minimisation Structurale du Risque) qui s'est montrée plus efficace que l'ERM (Empirical Risk Minimisation | Minimisation Empirique du Risque). Elle est notamment utilisée dans les réseaux neurones comme nous le reporte (Li et al., 2014), il reste tout de même des problèmes d'application à cause de la grande mémoire demandée,

pour y faire face différentes techniques ont vu le jours : le chunking (Osuna et al., 1997) ou le SMO (Sequential Minimal Optimization | Optimisation Séquentielle Minimale) (Osuna et al., 1997), pour lequel il suffit d'une matrice supplémentaire pour les calculs.

Le HMM représente l'approche la plus populaire. Elle se compose de trois étapes (Calliope, 1989) : la recherche de la probabilité que la séquence soit produite par le modèle et sa normalisation à travers l'algorithme de Viterbi ; la sélection des séquences capables de maximiser l'appartenance au modèle ; enfin un entraînement sur le modèle est effectué dans le but de maximiser la probabilité. L'approche par HMM montre ses limites dans certaines fréquences du signal et lorsque les traces sont bruitées, ce qui peut apporter une perte de données au moment de la modélisation (Mahola et al., 2007).

2.4 Les campagnes NIST

Le cadre d'évaluation NIST-SRE (National Institut of Standard and Technology-Speaker Recognition Evaluation) a su, au fil des années, fixer les chemins au niveau mondial pour le développement des systèmes et techniques de reconnaissance de la voix et du locuteur. Ces campagnes représentent un moment clé pour la recherche dans ce domaine, la dernière étant celle de 2016. Nous allons ici nommer les points intéressants et certains systèmes présentés lors de ces campagnes.

Les campagnes NIST se concentrent sur un ensemble de tâches de reconnaissance du signal sonore, cinq différents domaines y sont traités certains ayant apparu plus récemment. N'importe qui peut participer à ces campagnes et évaluer l'efficacité d'un système à travers les mesures établies par l'institution. Les deux campagnes les plus anciennes sont LRE (Language Recognition Evaluation | Évaluation de la Reconnaissance de la Langue) et SRE (Speaker Recognition Evaluation | Évaluation de la Reconnaissance du Locuteur), qui nous intéresse le plus et dont l'édition 2016, par rapport aux années précédentes, a apporté des différences importantes notamment : les données de résultats cette fois et comme en (NIST-SRE, 2012) ne seront pas fournis à l'avance ; un milieu contrôlé pour les enregistrements est introduit et une majeure variabilité de durée pour les données de test (de 10 à 60 secondes) ; des échantillons de moins de 9 secondes seront également utilisés pour étudier

le comportement des systèmes ; il n'y a pas de mélange de sexe ni de langue entre entraînement et test.

En 2010 le HASR (Human Assisted Speaker Recognition | Reconnaissance Assisté du Locuteur) a vu le jour dont une première expérience nous est rapportée par (Kahn, 2011), en 2013 comme déjà cité plus haut, l'apparition d'une première campagne pour l'évaluation et l'amélioration des algorithmes des i-vectors dans le domaine de la reconnaissance du locuteur lors des conversations téléphoniques a eu lieu.

La dernière campagne débutant en 2009 s'appelle Speaker Recognition for Biometrics, Forensics and for Investigatory purposes (Reconnaissance du Locuteur à des fins Biométriques, Judiciaires et d'Investigation). Elle représente un autre thème important dans le domaine de la reconnaissance du locuteur avec notamment son application dans le domaine judiciaire. En France ce thème a fait l'objet d'étude à plusieurs reprises par (Boë, 2000 ; Boë et Bonastre, 2012 ; Boë et al., 2018 ; Bonastre, 2003) qui invitent à être prudent sur l'actuelle fiabilité des systèmes, surtout lorsqu'une réponse erronée pourrait avoir des conséquences importantes. Notamment une motion (Société Française d'Acoustique, 1990) adopté en 1990 interdit l'utilisation de systèmes de RAL au sein des tribunaux tant que la recherche ne produira pas de résultats plus satisfaisants.

2.4.1 Systèmes extraits de NIST 2016

Parmi les systèmes participants à la campagne (NIST-SRE 16) nous avons voulu en sélectionner cinq qui représentent bien l'état de l'art, aussi bien par leurs réalisations que par leurs perspectives envisagées pour les futures améliorations du domaine.

Celui qui montre le taux d'erreur le moins élevé est décrit dans (Sadjadi et al., 2016b) avec un EER (Equal Error Rate | Taux d'Erreur Moyen) qui se maintient autour de 1.40%. Ce système est développé en fusionnant la modélisation du locuteur et du canal. La prise de décision est effectuée à travers des mesures linéaires. IBM confirme être capable de produire un EER très bas lors de ses participations (Sadjadi et al., 2016a). Le deuxième système choisi est décrit par (Shon, 2017), il utilise un mécanisme d'analyse assez atypique voire en contradiction avec les directives de la campagne 2016. Ce système partage l'idée de fond de (Dunbar et al., 2017) : entraînement et test sont effectués sur des langues

différentes. Pour compenser la difficulté de la tâche les auteurs fournissent des scores de normalisation sur ce même facteur, sur le genre et sur la variabilité inter linguistique, ce qui permet aux systèmes développés (5 au total utilisés de manière conjointe et indépendante) de garder un EER entre 13% et 18%. Ce taux semble être parfaitement dans la moyenne comme dans (Zhang et al., 2017a) ou (Rouvier et al., 2017), ce dernier regroupe les systèmes développés par le LIA (Laboratoire d'Informatique d'Avignon). Une série de systèmes utilisant les différentes techniques d'extraction de paramètres et de modélisation des locuteurs décrites plus haut, mais aussi des alternatives de fenêtrage et de filtrage du signal sonore, permettent de garder un EER moyen entre 15% et 18% et de baisser ce dernier à 14% avec l'utilisation d'un système hybride. Sur la même ligne se place (Khosravani et al., 2016) qui se concentre sur l'analyse des i-vector et nous affirme que l'approche par DNN (Deep Neural Networks | Réseaux Neurones Profonds) gagne en efficacité dans les segments appelés low SNR (Signal to Noise Ratio | Rapport Signal Bruit), c'est-à-dire lorsque le signal est bruité. Plus loin dans notre étude nous analyserons une mesure proche de celle-ci, le HNR (Harmonics to Noise Ratio | Rapport Harmoniques sur bruit).

2.5 Boîtes à outils

Le traitement du signal de parole ayant en informatique un rôle de plus en plus important et le développement de nouvelles techniques étant loin de s'arrêter, ce n'est pas étonnant que des APIs (Microsoft Corporation, 2018) pour la reconnaissance et la synthèse vocale soient présentes dans les applications Microsoft depuis le système Windows95. De la même façon, les systèmes d'exploitation Mac supportent nativement synthèse et reconnaissance depuis la version OS X 10.3 de 2003 (Apple Inc., 2018) (à travers l'introduction de la commande "say" sur terminal laquelle reproduit une synthèse du texte donné en paramètre) ou encore le logiciel libre eSpeak (Free Software Foundation, 2007) permettant d'avoir un système multilingue de Text To Speech sur tous les systèmes d'exploitation. Tous ces systèmes font partie d'un ensemble d'outils de synthèse vocale ou reconnaissance de la parole. Avec la récente apparition des assistants vocaux, la reconnaissance du locuteur a aussi pris sa place avec CORTANA et SIRI capables de répondre à une voix plus ou moins

neutre uniquement dans la langue d'entraînement de l'utilisateur dont le système a pu construire le modèle au préalable (à travers principalement des techniques de DNN comme rapporté dans les documentations respectives) (Microsoft Corporation, 2018 ; Apple Inc., 2017).

Du côté de la recherche, les outils permettant de faciliter la construction de systèmes et d'analyser leur efficacité ne manquent pas. Le toolkit Kaldi (Povey et al., 2011) paraît être le plus complet et le plus accepté par la communauté. Il a été retrouvé dans la plupart des travaux cités jusqu'ici. De la même façon, l'outil BOSARIS (Brümmer et de Villiers, 2010) représente une aide à la recherche pour la RAL largement utilisé et développé au sein de la société AGNITiO. Il propose des solutions de reconnaissance vocale et d'identification du locuteur dans le domaine judiciaire qui a été protagoniste en France à travers son logiciel BATVOX (basé sur i-vectors) utilisé par la Police Technique et Scientifique de Lyon lors d'une investigation en 2013 (AGNITiO, 2015).

3 Corpus de travail

Dans le cadre de notre étude aucun corpus spécifique n'a été créé. Pour retrouver à l'intérieur du signal de parole des marqueurs discriminants qui déterminent la variation interlocuteur et diminuent au maximum celle intralocuteur, deux corpus déjà existant ont été choisis pour effectuer les mesures souhaitées. Dans cette section une description des extraits qui ont été mis à notre disposition sera faite.

L'analyse des extraits sonores a été effectuée à travers le logiciels Praat (Boersma et Weenink, version 6.0.37) et certains calculs statistiques ont été obtenus à l'aide du logiciel R (R, version 3.4.4). Tous les scripts écrits et utilisés seront présentés dans le chapitre 5.

3.1 Corpus PATATRA

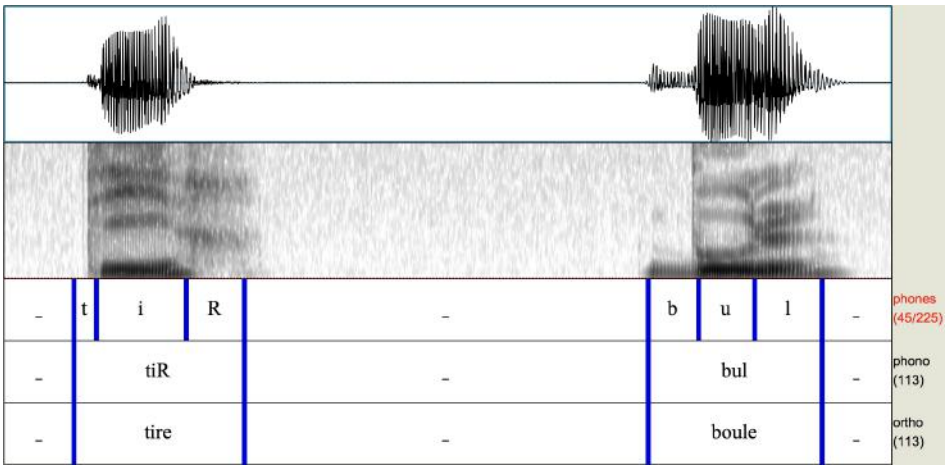

Fig. 1 Extrait de 2s mots lus PATATRA

Les premières données sont issues d'un corpus nommé PATATRA (Parole AdulTe À TRavers les Âges) récolté au sein du Laboratoire de Phonétique et Phonologie (LPP), créé dans le but d'étudier le vieillissement de la voix en recherchant des marqueurs de variabilité intralocuteur au fil du temps, à l'opposé de ceux que nous voulons identifier ici. Les enregistrements se composent de trois répétitions annuelles d'une liste de mots (entre 55 et 58), de la lecture du "la bise et le soleil", d'exercices de phonation et d'une séquence de parole spontanée. Nos mesures seront effectuées uniquement sur les liste de mots, à partir

desquelles nous pourrons en extraire voyelles et consonnes et les textes lus qui seront considérés aussi en tant que segment simple avec l'extraction d'une voyelle nasale.

Les enregistrements ayant débuté en 2013, 5 ans de répétitions sont aujourd'hui disponibles pour 10 locuteurs (4 hommes et 6 femmes), l'annotation, en Figure 1, est effectué manuellement sur Praat. Pour la transcription nous avons respecté les normes SAMPA (Wells, 1997). Les trois tires présentes dans les TextGrid[1] ont été un choix délibéré pour permettre une analyse plus précise et pour pouvoir s'intéresser également aux mots et aux segments isolés.

3.2 Corpus FABIOLE

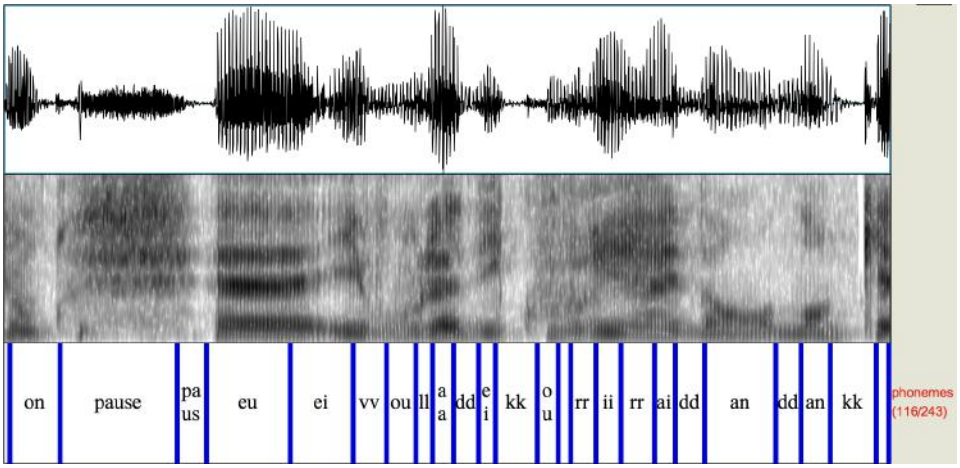

Fig. 2 Extrait de 2s FABIOLE

Les autres extraits auxquels nous avons eu accès appartiennent à la base FABIOLE. Il s'agit cette fois d'enregistrements issus d'émissions radios et télévisées, le milieu n'est donc pas contrôlé cependant la qualité audio est assez bonne ce qui est visible à travers le spectrogramme de la Figure 2. La durée moyenne est de 30 secondes par session (cependant les enregistrements ne sont pas forcément consécutifs ils sont parfois fragmentés en plusieurs fichiers), un minimum de 30 sessions pour 100 locuteurs sont disponibles. En s'agissant ici de voix journalistiques nous observerons sans doute des différences importantes lors de l'application des mesures de rythme et de fréquence fondamentale comme suggéré dans (Gendrot et al., 2012). Nous allons voir dans le chapitre

[1] un des types d'objet de Praat désignant les annotations qui peuvent être associées à un fichier audio ou bien être analysées singulièrement.

suivant quels sont les indices étudiés et lors de la discussion en quelles mesures leur variation sera pertinente pour l'un ou l'autre ensemble de données. L'annotation a été effectuée en partie automatiquement, ce qui peut expliquer le majeur point faible de cette base : une forte imprécision des segmentations. Ne s'appuyant pas sur les règles SAMPA une légère adaptation des scripts pour l'extraction des phonèmes nous intéressant a été nécessaire, de plus la présence d'une seule tire dans les TextGrid de cette base nous limite au travail par segments isolés sur l'ensemble du signal sonore, nous empêchant d'analyser pour le moment les mots entiers ou d'autres unités qui en revanche peuvent être facilement retrouvés dans la base décrite précédemment. Cela représente peut-être une des limites de ce corpus et un des points à améliorer pour les prochaines distributions. L'ensemble de la base FABIOLE et ses applications ainsi que l'évolution qu'elle pourra avoir dans le futur ont été décrites de manière détaillée dans (Ajili et al., 2016b).

4 Méthodes

Nous allons présenter dans cette section les mesures acoustiques qui ont été étudiées lors de l'analyse. Nous expliquerons leur sélection ainsi que leur impact en les comparant aux mesures habituellement utilisées dans le domaine de la RAL et du traitement du signal. La présente étude se positionne dans la même lignée que les travaux de (Kahn, 2011), le point de départ posé par l'auteur est notamment la recherche de nouveaux marqueurs de variation interlocuteur pertinents qui soient suffisamment constants au niveau intralocuteur. Dans notre étude ils comprennent : l'analyse des moments spectraux, des mesures sur la fréquence fondamentale et sur le rythme ainsi que la prise en considération des MFCC.

Pour obtenir tous ces éléments à partir des deux corpus considérés un script Praat a été mis en œuvre, les différences entre chaque base sont minimes. Ces mesures ont été choisies pour leur représentation à la fois des propriétés statiques appartenant aux informations acoustiques présentes dans le signal sonore et dynamiques qui expliquent les changements et liens qui existent entre ces informations à l'intérieur d'un signal sonore.

4.1.1 Analyse de variance et taille d'effet

Les mesures citées et présentées dans les prochains paragraphes, représentent les variables que nous testons afin d'obtenir un élément de discrimination qui puisse montrer sa cohérence au niveau phonémique intralocuteur. Nous allons exploiter au maximum les résultats obtenus pour comprendre leur fiabilité et pour faire ceci nous avons besoin d'une mesure cohérente, elle-même fiable au sein de l'évaluation, (Kahn, 2011) utilise le calcul de la taille de l'effet du facteur locuteur sous forme d'êta carré (notée η^2) dans son étude. Pour pouvoir comparer nos résultats nous ferons de même. Une fois toutes les mesures calculées, une analyse de variance (ANOVA) prenant le locuteur comme facteur fixe et le marquer étudié comme variable dépendante est effectuée. À partir de ces résultats la taille de l'effet de chaque indice sera calculée, la mesure d'êta carré est introduite par (Levine et Hullett, 2002) pour expliquer l'influence d'un facteur fixe lors d'une analyse de variance. Les auteurs décrivent plusieurs propriétés majoritairement liées aux propriétés calculatoires de la taille d'effet par le biais de l'êta carré (la simplicité du calcul, la somme totale des

composantes égale à 1.00, etc.) en pointant aussi sur le fait qu'il est toujours inférieur ou égal à l'êta carré partiel et conséquemment perçu comme plus conservateur. La caractéristique qui nous intéresse le plus est la représentation de l'êta en pourcentage, en multipliant sa valeur par 100 et non pas selon un seuil, comme pour le p-value.

4.1.2 Analyse en composantes principales

Dans un deuxième temps nous allons nous intéresser à un niveau plus élevé de la variabilité inter locuteur, à la place de considérer le niveau interphonémique pour la recherche d'indices nous projetterons les locuteurs dans un espace. Ceci nous permettra d'observer les distances entre eux et comprendre en quelle mesure les indices que nous avons calculés sur le signal sonore ont une incidence réelle sur la distinction des locuteurs. Pour pouvoir effectuer cette nouvelle représentation nous allons nous servir de l'Analyse en Composantes Principales (ACP) décrite pour la première fois par (Pearson, 1901), une technique répandue en analyse de données (Jolliffe, 1986 ; Wold et al., 1987) surtout lorsque celles-ci se trouvent être d'une dimension conséquente (Besse, 1992).

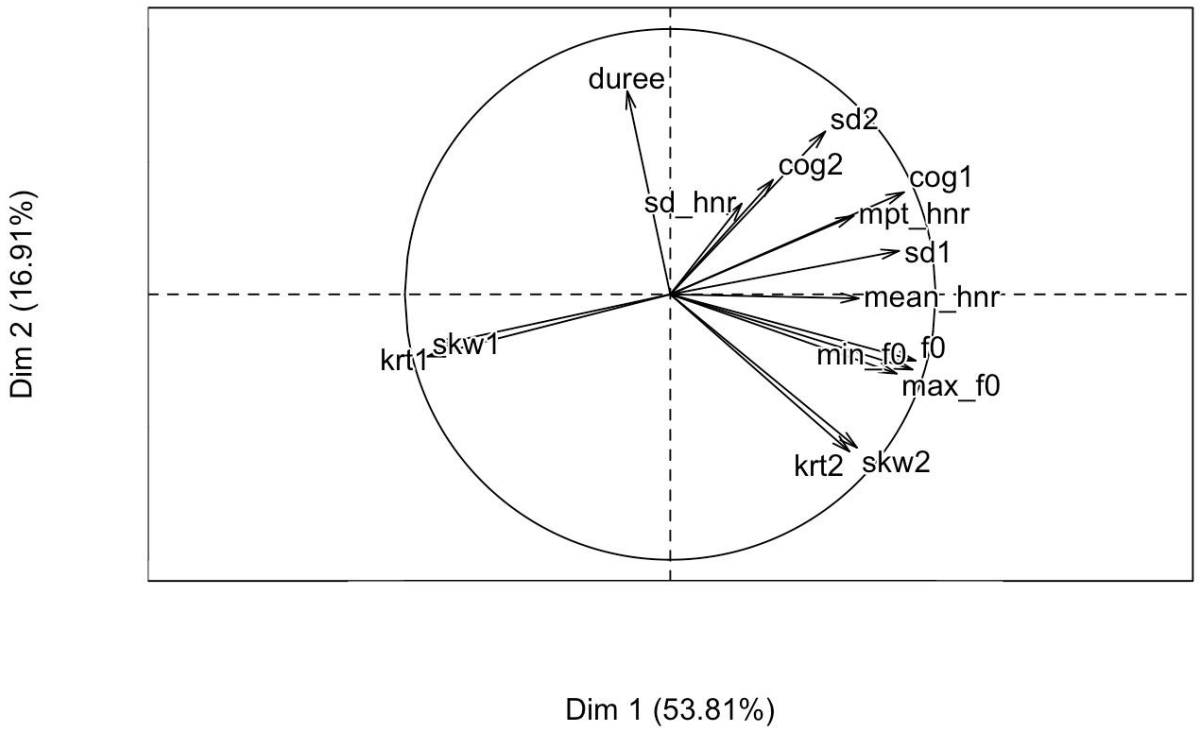

Fig. 3 Nuage des variables ACP PATATRA, données normalisées

Nous allons brièvement présenter les principes de cette méthode de façon à avoir une idée générale des concepts qui seront plus développés lors de la discussion. L'ACP consiste à réduire la dimension du problème examiné en extrayant les informations essentielles grâce à une projection dans un espace de variables plus petit. Dans notre cas, ce qui est appelé problème est le tableau des observations pour chaque phonème de chaque locuteur alors que les variables sont représentées par les différents indices observés. Ainsi si dans notre jeu de données la représentation des locuteurs est effectuée grâce à l'ensemble de ces variables, lorsque nous nous retrouvons dans un espace réduit ces variables deviennent indépendantes et sont projetées sur les composantes ou axes à travers une fonction de cosinus comme le montre la Figure 3 ci-dessous. La position de ces éléments dans l'espace n'influencera plus la représentation des individus, les locuteurs dans notre cas, par rapport aux axes. Elle nous fournira une information supplémentaire sur le rôle que jouent nos variables sur les individus, nous montrerons ceci dans les représentations suivantes.

Un autre concept à retenir lors d'une Analyse par Composantes Principales est le pourcentage d'inertie, une valeur qui permet de quantifier l'information des composantes. Lorsque les nouveaux axes sont construits nous gagnons en représentation : nous avons une image plus claire de la distribution des individus, de la dispersion de la population observée et une vision plus facilement interprétable. Nous risquons cependant de perdre de l'information car à travers les composantes le nombre de variable est réduit et les informations ne sont plus corrélées aux observations de départ. Les Composantes Principales étant les résultats de la matrice de covariance de nos données leur dispersion, appelée inertie (Sarkar et al., 2014), nous permet d'identifier le pourcentage d'information qu'elles expliquent et en conséquence de choisir quels axes seront les plus représentatifs pour notre espace. Dans l'exemple ci-dessus nous voyons que les deux premières composantes portent respectivement 53.81% et 16.91% de l'information, nous pourrons donc décrire la variabilité de la population à 70.72%, ce qui représente le taux le plus élevé pour ce premier ensemble de données.

Tous les calculs d'ACP ont été effectués à travers les fonctionnalités présentes par défaut dans le logiciel R (R, version 3.4.4) et dans la librairie FactoMineR (Josse et Husson, 2008).

4.1.3 Arbres de décision

L'autre méthode, dont les résultats sur les distances entre locuteurs seront vus et discutés ultérieurement, est celle des arbres binaires de décision plus généralement provenant des modèles dits CART (Classification and Regression Tree) (Breiman et al., 1984 ; Quinlan, 1986) et que nous avons obtenus à travers le logiciel R (R, version 3.4.4). Il s'agira également de rendre plus lisible et facilement interprétable les données à notre disposition, comme lors de l'ACP, en gardant cette fois les variables de départ, aucune variable ou composante supplémentaire sera ajoutée à la représentation de la population.

Un arbre de décision est une représentation graphique des informations du corpus et des décisions prises pour classifier les individus selon un ensemble homogène de variables, les locuteurs et leurs indices acoustiques correspondant dans notre cas. Toutes les observations sont regroupées à la racine de l'arbre, une première variable avec une réponse binaire est prise en compte pour donner lieu à deux branches divisant l'ensemble des données. Les nœuds qui suivront pourront donner d'autres branches au fur et à mesure que la classification se complexifie ou se résoudre dans des extrémités ou feuilles. Un ensemble peu homogène de données se traduira en un arbre très complexe car ces modèles de représentation sont très sensibles à la fluctuation des échantillons. Ceci est l'inconvénient majeur de cette représentation.

4.2.1 Moments Spectraux

Le premier ensemble de mesures que nous avons décidé d'examiner correspond aux moments spectraux (MS). Il s'agit d'une série de calculs obtenus à travers la puissance du module du spectre multiplié par pondération du domaine fréquentiel : la pondération sera effectuée selon le spectre[2] lui-même si la puissance est égale à 2 ; selon la valeur absolue du spectre si la puissance est de 1 (Boyer et al., 2001). Nous avons utilisé ces deux valeurs de la puissance pour nos calculs afin d'observer si cela donne lieu à des différences remarquables dans les résultats. Les moments spectraux représentent avec d'autres mesures comme l'amplitude, la durée du bruit ou encore les point de pics spectraux des

[2] un spectre est le résultat de l'application de la transformée de Fourier sur un signal sonore.

propriétés statiques du signal sonore. L'analyse des moments spectraux est une procédure statistique qui considère à la fois la moyenne de la fréquence locale et globale du spectre ainsi que les aspects de variabilité du signal vocal. Nous nous baserons sur la littérature existante en phonétique, en particulier (Jongman et al., 2000 ; Spinu et Lilley, 2016), pour expliquer la valeur des MS et pourquoi nous avons décidé de les utiliser. Le traitement du signal vocal n'est pas le seul domaine dans lequel les moments spectraux ont du succès, ils existent de nombreux travaux sur leur utilisation. (Boyer et al., 2001) ont validé leur utilité en physique pour le traitement des signaux photo acoustiques, se basant aussi sur l'analyse d'un spectre. Ce n'est pas étonnant que les considérations faites par ces auteurs se rapprochent de celles qui concernent plus notre domaine.

Le principe à partir duquel nous avons décidé d'appliquer l'analyse par moments spectraux à la reconnaissance du locuteur réside dans le fait qu'il existe plusieurs moments chacun résultant du calcul appliqué et renvoyant à une caractéristique particulière du spectre. Par conséquent, les moments spectraux étant une représentation de ce dernier sous une autre dimension, si nous arrivons à identifier un moment précis caractérisant plus que les autres un locuteur alors nous pourrons trouver une mesure fiable pour la discrimination. L'analyse par moments spectraux a connu son succès dans le traitement du signal sonore surtout pour sa bonne réussite dans la classification des occlusives et des fricatives avec les deuxième et troisième moments considérés comme les majeurs contributeurs dans la décision et le premier qui est le plus porteur d'informations (Spinu et Lilley, 2016). Ceci est dû à la forme générale d'un spectre, en prenant l'exemple d'une fricative, celle-ci est déterminée par la dimension et la forme de la cavité devant la constriction, plus longue sera cette cavité plus définie sera la forme du spectre. Les fricatives alvéolaires et palato-alvéolaires seront caractérisées par des spectres bien définis alors que les fricatives labiodentale et interdentales montrent un spectre plat (Martin, 2008).

Les deux premiers moments, le centre de gravité (Spectrum: Get center of gravity sur Praat) et la Variance ou écart type (Spectrum: Get standard deviation sur Praat) reflètent respectivement la concentration et l'étendue moyenne de l'énergie. Le troisième, le coefficient d'asymétrie (Spectrum: Get skewness sur Praat) comme son nom le suggère est un indicateur de l'asymétrie de la distribution de l'énergie dans le spectre. Un coefficient égal à 0 indique une distribution symétrique autour de la moyenne, ainsi elle est positive quand la

partie droite de la distribution est plus étendue que celle de gauche, de la même façon elle est négative quand la partie gauche de la distribution s'étend plus que celle de droite. Du point de vue purement phonétique, l'asymétrie de la distribution est le résultat de la concentration de l'énergie, une valeur positive correspondra à une prédominance de l'énergie dans les basse fréquences, une négative à une prédominance dans les hautes fréquences (Jongman et al., 2000). Le quatrième moment correspond au coefficient d'aplatissement (Spectrum: Get kurtosis sur Praat), il mesure directement l'acuité du pic de la distribution, si sa valeur est positive un pic plus important sera présent autour de la moyenne. Lorsque cette mesure est appliquée au spectre un kurtosis positif correspondra à une forme plus définie avec des pics bien formés, alors qu'avec un coefficient négatif le spectre sera plutôt plat.

En conclusion, les moments spectraux peuvent être calculés sur deux échelles, Bark et linéaire, les deux n'étant pas très différentes. Cependant la première est considérée comme légèrement plus efficace lors des analyses (Spinu et Lilley, 2016).

4.2.2 Rapport Harmoniques sur bruit

Lors de l'écoute l'humain arrive, sans vraiment s'en rendre compte, à distinguer des composantes harmoniques, provenant des vibrations laryngiennes, et des composantes de bruit, qui peuvent être associées aux frictions, occlusions, clics mais aussi à d'autres facteurs, jitter, shimmer, désordres des plis vocaux (Martin, 2008). Ces deux composantes étant présentes dans le signal sonore, pour l'analyse automatique d'un signal il paraît donc très utile de réussir à faire la même distinction. En séparant les deux composantes le rapport entre leur énergies peut être facilement calculé, le résultat est appelé HNR et fournit un indicateur de la périodicité acoustique globale du signal vocal en quantifiant le rapport entre les composantes périodiques (partie harmonique) et apériodiques (le bruit) (Yumoto et al., 1982).

Ce qui fait principalement la distinction entre les harmonies et le bruit est la structure périodique de l'une qui ne se retrouve pas dans la deuxième. Cette périodicité, ou plutôt semi-périodicité puisque nous n'avons pas à faire à des sons purs (Martin, 2008), est due au fait que les harmonies présentes dans le signal sonore proviennent de la répétition d'une

fréquence de base, appelée fréquence fondamentale et notée généralement f_0 (Calliope, 1989). La fréquence fondamentale est un autre des éléments qui sera pris en considération pour rechercher un indice fiable de la variabilité interlocuteur car elle est fortement liée à la source, la taille des cavités et leurs formes, qui l'influencent de manière considérable (Vaissière, 2006). Nous allons considérer les valeurs moyennes, minimales et maximales ainsi que son étendue sur les segments présents dans nos corpus.

Reprenons maintenant le HNR, (Oller, 2008 ; Yumoto et al., 1982) nous rapportent comment son utilisation est strictement liée au champ médical pour comprendre le comportement des plis vocaux. Comme pour jitter et shimmer dans (Kahn, 2011) ici nous essayons de retrouver des indices généraux de variabilité inter locuteur avec une mesure habituellement utilisée pour diagnostiquer des désordres à l'intérieur du signal vocal. Tous les outils de diagnostic de la voix qui calculent ce paramètre utilisant une approche sur la fréquence se maintiennent en dessous de 5 kHz. Les mêmes auteurs affirment qu'il n'existe pas un moyen unique de calculer ce rapport et les méthodes pour l'obtenir peuvent être classifiées en deux types : basée sur la fréquence ou le temps. Le calcul pourrait être effectué en se basant sur le cepstre mais pour pouvoir exploiter son résultat une estimation basée sur la fréquence serait nécessaire, reliant cette autre méthode à celles du premier type. Pour notre étude nous nous sommes basés sur la méthode qui est présentée par Praat et dont nous reportons ici l'entrée du manuel avec quelques commentaires. Le HNR, exprimant un rapport entre deux dimensions utilise le décibel (dB) comme unité de grandeur : si 99% de l'énergie du signal se retrouve dans la composante périodique et le bruit ne possède que le 1% restant alors le HNR sera de 10*log10(99/1) = 20dB. Un HNR de 0dB montre que l'énergie du signal est divisée de manière égalitaire entre les harmonies et le bruit. En moyenne un locuteur ne présentant pas de désordres des plis vocaux pourra produire un /a/ avec un rapport harmoniques sur bruit autour de 20dB et un /u/ avec un HNR de 40dB.

4.2.3 Les mesures de rythme

Les mesures de rythme que nous avons décidé de considérer dans notre recherche sur la variabilité interlocuteur se divisent principalement en deux familles : des mesures brutes et

des mesures normalisées (Gharsellaoui et al., 2018). Les deux effectuent un calcul sur le discours en segmentant le signal vocal en intervalles vocaliques et consonantiques. Le PVI (Pairwise Variability Index | Indice de Variabilité Apparié) mesure la différence de durée entre des intervalles immédiatement consécutifs et la moyenne entre ces différences sur toute la réalisation.

Le PVI normalisé (nPVI) est une mesure de la variation moyenne d'un ensemble de distances (durées) qui sont obtenues à partir de paires successives adjacentes d'éléments (Nolan et Asu, 2009). Il était au début créé pour calculer les différences rythmiques entre les langue sur la base de la longueur des voyelles et plusieurs applications dans ce sens ont été réalisées avec des résultats intéressants (Gharsellaoui et al., 2018). Les recherches de (Nolan et Asu, 2009) font une revue de l'histoire et des applications du PVI dans les sciences du langage. Le nPVI s'est montré être l'indice le plus approprié dans l'identification de l'estonien parmi les langues étudiées, il a également été appliqué pour déterminer la complexité cognitive d'un texte. Plus récemment il a été suggéré, notamment dans (Toussaint, 2013), que le nPVI devienne un outil utilisé dans l'analyse du rythme musical, les résultats montrés confirme les bonnes performances dans la distinction de styles musicaux en français et allemand ou dans des tests en lien avec le domaine psychologique. Le nPVI a porté à l'évidence l'existence d'un code commun de l'expression d'émotion dans le discours et en musique par le rythme du discours.

Dans (Gharsellaoui et al., 2018 ; Nolan et Asu, 2009 ; Toussaint, 2013) les auteurs tendent à utiliser les résultats bruts, noté rPVI, lorsqu'ils mesurent l'effet sur les consonnes (nous l'indiquerons comme CrPVI) mais ils utilisent plutôt des mesures normalisées, nPVI, pour les voyelles (VnPVI). Les auteurs ont remarqué que l'utilisation des métriques normalisées améliore la discrimination entre différent types de voix. En utilisant le nPVI, résultant des intervalles vocaliques, une meilleure performance dans la discrimination des langues est obtenue plutôt qu'avec des mesures consonantiques ou non normalisées. (Toussaint, 2013) recommande la combinaison d'au moins deux mesures en proposant celle entre CrPVI et VnPVI comme la plus discriminante. Nous verrons par la suite quelle combinaison apporte le plus de résultats pour le domaine de la reconnaissance du locuteur.

Le calcul des mesures de PVI a été effectué à travers l'outils pvicalc (Trouville, 2016) avec des modifications mineures pour le réadapter aux besoins de cette étude.

4.2.4 MFCC

Le dernier ensemble de mesures que nous avons décidé de considérer lors de cette étude est un choix en quelque sorte obligé, selon ce que nous avons affirmé dans le chapitre décrivant l'état de l'art, pour pouvoir se confronter au domaine de la reconnaissance du locuteur. Il s'agit des Coefficients Cepstraux à Fréquence Mel, MFCC, qui ont déjà été abordés dans la partie dédiée aux méthodes de modélisation du locuteur habituellement retrouvées en RAL. Ces coefficients sont calculés à partir du cepstre du signal sonore qui est obtenu en appliquant, pour la deuxième fois, la transformée de Fourier au logarithme naturel de la transformée de Fourier du signal d'origine (Calliope, 1989).

Dans la littérature en phonétique acoustique, en moyenne un ensemble de 13 MFCC a été largement utilisé et avec succès pour la classification des fricatives (Davis et Mermelstein, 1980 ; Kamruzzaman et al., 2010), 85% de taux moyen de réussite, en anglais britannique et portugais (Jongman et al., 2000), avec un intérêt particulier porté au voisement. En comparant entre MFCC et SM (Spinu et Lilley, 2016), les premiers montrent un avantage particulier dans la classification des labiales lorsqu'elles sont considérées en opposition à d'autres consonnes et dans la classification du genre du locuteur, en particulier lorsqu'il s'agit d'analyser des segments consonantiques.

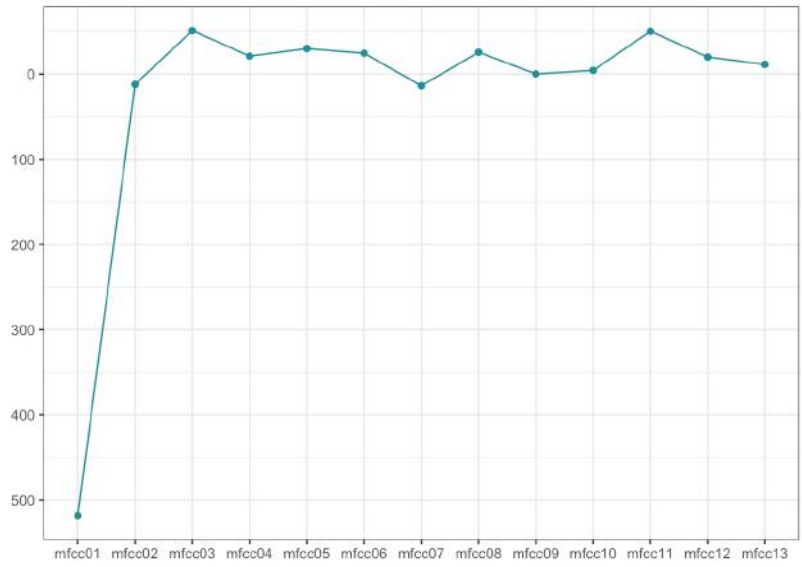

Fig. 4 MFCC pris au milieu d'une voyelle

Cette méthode a également montré une estimation très efficace des distances acoustiques parmi 13 différents accents des îles anglaises (Ferragne et Pellegrino, 2010), avec les auteurs qui arrivent à affirmer que “les MFCC ne peuvent pas se tromper (alors que les formants oui) [ceci] fournit un grand support à leur utilisation dans les études phonétiques”. Différemment des autres mesures décrites jusqu’ici, les MFCC fournissent une représentation dynamique du signal, comme le montre la Figure 4 ci-dessus, puisqu’ils reproduisent ce dernier selon une échelle différente (Davis et Mermelstein, 1980). Cependant pour pouvoir les évaluer dans le sens de notre travail et en comparer l’efficacité avec les autres indices acoustiques qui ont été choisis, nous prendrons chacun des 13 coefficients individuellement et les analyserons de manière statique.

5 Mesures acoustiques

Dans cette section nous allons présenter les résultats obtenus après l'application des différentes mesures ainsi que l'importance de leur effet, nous examinerons leurs points faibles et les informations qu'elles ont pu apporter. Le corpus PATATRA présentant un ensemble plus restreint de phonèmes, nous avons décidé de garder ceux qui sont présents dans cette base comme référence pour notre analyse[3] et permettre une comparaisons entre les valeurs qu'ils obtiennent pour les deux corpus. Nous retrouvons notamment pour les consonnes : les occlusives /p/ /t/ /k/ /b/ /d/, les fricatives /f/ /v/ /ʃ/ /z/, les liquides /l/ /ʁ/ et la nasale /m/. Pour les voyelles nous retrouvons une nasale /ã/ et les voyelles cardinales du français : deux voyelles fermées, une postérieure /i/, une arrière /u/ et une ouverte /a/. Ce même ensemble de segments a été extrait de la base FABIOLE, nous avons aussi gardé deux voyelles nasales supplémentaires qui sont ici présentes vu le bon classement de ces segments dans l'échelle de discrimination exposée plus haut (voir page 12). Nous allons maintenir pour ces deux segments les notations non conventionnelles /ɛ̃/ et /ɔ̃/ utilisées dans le corpus FABIOLE. Les résultats du premier corpus seront aussi discutés par sexe puisque nous avons un ensemble de 6 locutrices et 4 locuteurs contrairement à la base FABIOLE, qui recueille 30 locuteurs uniquement de sexe masculin.

Les valeurs d'êta carré seront présentées sous forme de pourcentage, les premiers deux chiffres décimales sont utilisés pour faire la correspondance avec la valeur réelle du η^2 (Richardson, 2011), ceci permettra de comprendre quelle influence a le facteur locuteur sur l'indice considéré. À partir de la valeur obtenue nous pourrons définir quelle est la taille d'effet pour l'indice pris en compte, toujours selon ce qui est reporté dans (Richardson, 2011) une taille d'effet de 0.0x est considérée comme une taille minime, à partir de 0.20 elle sera considérée moyenne alors qu'une importante taille d'effet se manifeste à partir de 0.60 puisqu'elle indique un effet d'au moins 60%.

Nous effectuerons aussi une comparaison entre les effets obtenus par nos mesures et ceux présentés par (Kahn, 2011) pour ainsi définir si les paramètres que nous avons choisi ont une meilleure performance par rapport à ceux qui ont déjà été étudiés. Le premier calcul

[3] Les tableaux complets reportant aussi les segments exclus lors des résultats sont présentés en annexe.

qui a été effectué sur toutes les mesures est une analyse de variance où le locuteur ne constitue pas le facteur fixe à analyser mais plutôt une variable, le facteur fixe étant l'année (dans PATATRA) et la session d'enregistrement, nous avons pu tester à quel point ces mesures sont influencées du point de vue intra locuteur. Les résultats obtenus sont assez satisfaisant de ce point de vue puisque les êta carré restent autour de 0%, d'un enregistrement à l'autre les indices paraissent constant dans la représentation d'un locuteur.

5.1 Moments spectraux

Chaque moment spectral a été calculé avec les deux valeurs en puissance, 1 et 2, exposées dans le chapitre précédent, nous allons reporter toutes les valeurs d'êta carré qui sont ensuite obtenues ceci permettra d'évaluer si un calcul est plus efficace que l'autre. Nous avons décidé d'effectuer les deux calculs parce que dans la littérature (Jongman et al., 2000 ; Boyer et al., 2001) et dans la documentation présente dans Praat sur le sujet ces deux valeurs sont utilisées le plus souvent pour montrer l'impact sur les résultats. Une troisième valeur, $^2/_3$, est conseillée dans Praat mais nous avons préféré l'exclure de notre étude actuelle.

p	**t**	**k**	**b**	**d**	**l**	**m**	**ʁ**	**f**	**ʃ**	**v**	**z**
13 %	4 %	7 %	14 %	27 %	45 %	46 %	32 %	43 %	41 %	28 %	46 %
ã	**a**	**i**	**u**								
61 %	55 %	47 %	40 %								
p	**t**	**k**	**b**	**d**	**l**	**m**	**ʁ**	**f**	**ʃ**	**v**	**z**
14 %	12 %	6 %	21 %	7 %	39 %	71 %	37 %	35 %	47 %	10 %	26 %
ã	**a**	**i**	**u**								
50 %	42 %	14 %	47 %								

Tableau 1 Pourcentage de la taille d'effet selon le η^2 pour le centre de gravité spectral - PATATRA

Pour le corpus PATATRA nous avons les valeurs globales d'êta carré du premier moment spectral (centre de gravité du spectre) qui restent en dessous de 30% pour les occlusives, avec l'incidence mineure pour /t/ et /k/ dans le premier calcul et pour /k/ et /d/ dans le deuxième cas. Les autres segments montrent un effet moyen (en gardant l'indication donné

plus haut de 0.60 comme valeur importante de la taille d'effet) avec seule la voyelle nasale qui arrive à montrer une taille d'effet considérable dans le premier calcul. Le deuxième cas du centre de gravité spectral montre aussi une détérioration des résultats pour la majorité des segments, /v/ passe de 0.28 à 0.10, /i/ passe de 0.47 à 0.14 mais dans le cas de /m/ nous observons plutôt une croissance de η^2 allant de 0.46 à 0.71.

En prenant séparément les locuteurs et les locutrices nous observons que les occlusives gardent des valeurs très basses /t/ 1% pour les hommes /k/ 5% dans les deux groupes, /d/ est le seul segment de cette classe qui dépasse 20% d'effet dans le cas du premier MS, uniquement les deux voyelles ouvertes, orale et nasale, dépassent le 0.50 de η^2, cependant aucun des segments considérés ne s'approche de la valeur de 0.60 de taille d'effet. Pour le corpus FABIOLE, ces premières valeurs de η^2 sont encore inférieures, les segments nasals et la voyelle orale /a/ dépassent à peine 30% de taille d'effet. Les occlusives restent sur des valeurs très peu discriminantes et les fricatives montrent une réussite bien inférieure par rapport aux données de l'autre corpus considéré, ceci nous amène à considérer le premier moment spectral comme une mesure peu déterminante.

p	**t**	**k**	**b**	**d**	**l**	**m**	**ʁ**	**f**	**ʃ**	**v**	**z**
6 %	2 %	1 %	20 %	14 %	14 %	31 %	8 %	8 %	15 %	13 %	9 %
ɑ̃	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**						
33 %	37 %	25 %	30 %	14 %	9 %						
p	**t**	**k**	**b**	**d**	**l**	**m**	**ʁ**	**f**	**ʃ**	**v**	**z**
2 %	1 %	1 %	17 %	15 %	14 %	19 %	6 %	8 %	10 %	11 %	7 %
ɑ̃	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**						
27 %	32 %	16 %	27 %	10 %	2 %						

Tableau 2 Pourcentage de la taille d'effet selon le η^2 pour le centre de gravité spectral - FABIOLE

Le premier moment spectral reste cependant celui qui montre les résultats les plus intéressants parmi les quatre. Les liquides et les fricatives, au moins pour le corpus PATATRA, montrent des effets supérieures aux autres segments avec les nasales qui représentent des facteurs peu discriminant certes mais qui pourraient faire l'objet d'analyses supplémentaires pour exploiter au mieux leur potentiel. Les valeurs des tailles d'effet des

autres moments oscillent entre 0.40 et 0.20 pour la majorité des segments. Les fricatives non voisées montrent des valeurs plus élevées dans les premiers calculs avec leur effet qui se réduit de moitié dans la deuxième évaluation. Les nasales continuent à être des segments montrant une taille d'effet moyenne, /m/ est la consonne avec les scores les plus élevés après les fricatives non voisées. Les voyelles nasales que nous avions à disposition dans les deux corpus ont un effet minime de 50% dans le corpus PATATRA et de 30% pour le corpus FABIOLE. Pour les moments de skewness et kurtosis, considérés uniquement pour les locuteurs hommes du corpus PATATRA nous obtenons avec la voyelle nasale /ã/ l'effet discriminant majeur pour la mesure des moments spectraux dans notre analyse avec une valeur de 62% et 60%. Ce même segment pour les locutrices présente un effet de 58% et 57%. Dans le corpus FABIOLE les deux mesures que nous venons de citer montrent un effet d'uniquement 20% pour les segments nasals, cette différence de performance si élevée peut être expliquée par divers facteurs : d'un côté, elle montre l'écart d'analyse entre des enregistrements provenant d'un milieu contrôlé, la base PATATRA notamment, présentant des mots et du texte lus en chambre sourde et ceux de la base FABIOLE extraits d'émissions télévisées et radiophoniques ; d'un autre côté cet écart met l'accent sur le problème de la segmentation, si la base FABIOLE nous offre un grand nombre de données permettant une analyses sur des quantités considérables (trois fois plus de locuteurs, 10 fois plus long en terme de durée d'enregistrements), ces données restent annotées automatiquement, nous ne saurons pas avec exactitude si toutes les segmentations des phonèmes et les annotations sont correctes.

5.2 HNR

Sur le Rapport Harmoniques sur bruit nous avons appliqué trois calculs afin de comprendre si les valeurs pour chaque segment sont plus ou moins indicatifs : médiane, moyenne et écart type. Les résultats de la médiane n'ont pas d'effet significatif, si ce n'est pour la voyelle nasale de PATATRA avec 54% et les segments /m/, /ʃ/, /a/, /u/, pour lesquels des valeurs de 24% à 28% sont obtenues, uniquement 6 dépassent le 5% d'effet sur le facteur locuteur. L'écart type arrive à des valeurs bien en dessous de la médiane avec aucun

segment dépassant le 0.20 de η^2. En ce qui concerne la moyenne des valeurs légèrement plus significatives sont observées mais uniquement à l'intérieur du corpus PATATRA.

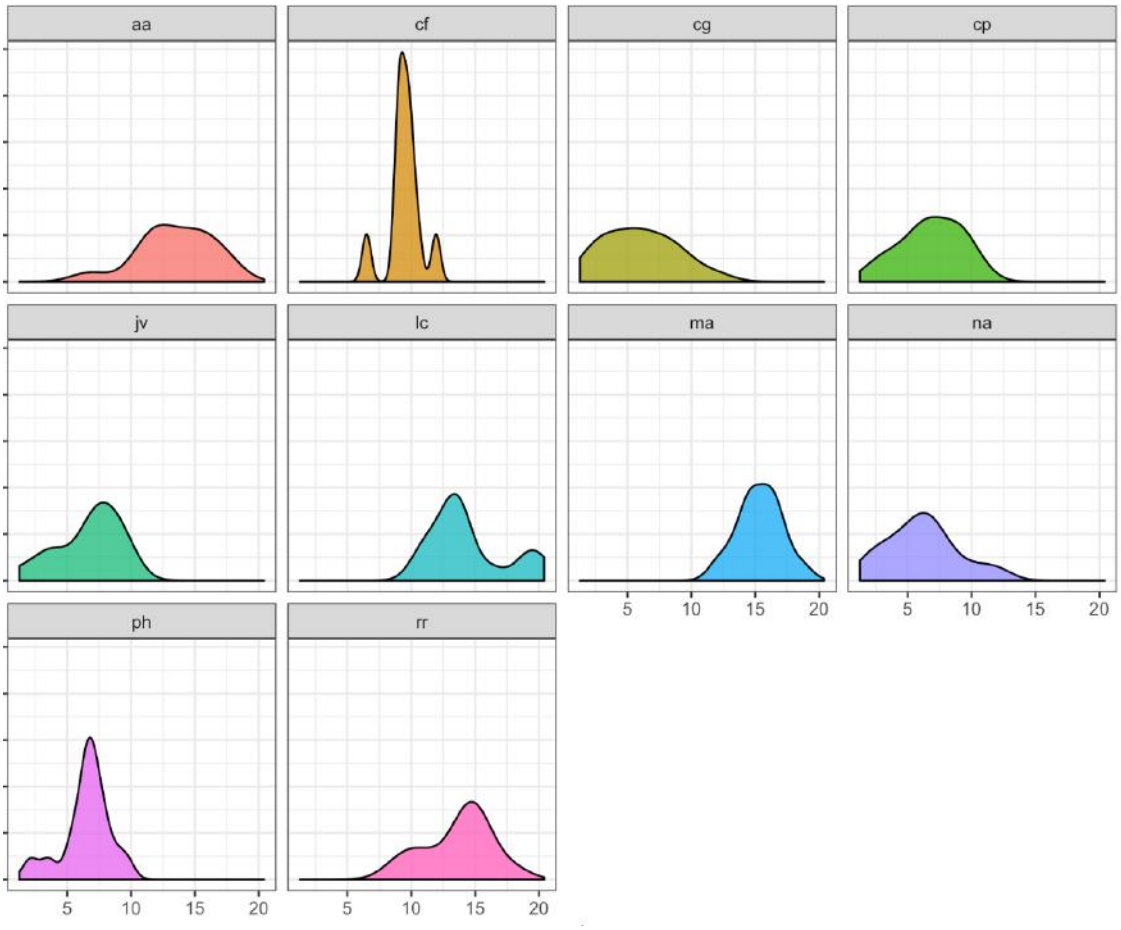

Fig. 5 valeurs de HNR (dB) par locuteur pour la nasale /ã/ dans le corpus PATATRA sous forme de densité spectrale

La voyelle nasale /ã/ dépasse la valeur de 0.60 d'êta carré en mesure globale, lorsque uniquement l'ensemble des locuteurs hommes est considéré cette valeur s'élève à 70% d'effet et elle reste invariée à 61% pour les locutrices. Ceci se traduit dans la Figure 5 ci-dessus par des représentations sous forme de densité spectrale variables pour les différents locuteurs, chaque motif décrit les valeurs de HNR des dix locuteurs du corpus PATATRA pour ladite voyelle nasale en gardant les informations liées au spectre. La position de la densité dans cette représentation indique la dimension de la valeur comme indiqué par l'axe des abscisses : un locuteur qui sera placé à gauche, jv et ph, les deux premiers des deuxième et troisième lignes par exemple, présentera des valeurs moins élevées qu'un locuteur placé vers la droite, ma par exemple, troisième motif de la deuxième ligne. Plus la

densité aura un spectre aigu plus le nombre d'observations pour cette valeurs sera important, c'est ce que nous pouvons constater pour deux locuteurs qui ne présentent pas de valeurs très élevées mais en nombre conséquent. Nous le rappelons le HNR est une mesure utilisée surtout pour étudier les comportements des plis vocaux, nos locuteurs ne présentant pas de dysfonctionnements à ces organes cet indice n'est pas un élément représentatif de la variabilité inter locuteur. Cette mesure se montre plus efficace que les indices de jitter et shimmer dans (Kahn, 2011), lesquels sont aussi utilisés dans des contextes médicaux, cependant avec le rapport Harmoniques sur bruit les résultats attendus ne sont pas obtenus et il ne peut pas être considéré comme un facteur déterminant pour le repérage d'éléments propre au locuteur à l'intérieur du signal vocal. Nous observons un effet sur la seule voyelle nasale à notre disposition dans ce corpus, il pourrait être étudié de manière plus approfondi dans un contexte de voix contrôlée pour comprendre si avec des échantillons plus nombreux à la fois en terme de locuteurs, 9 dans ce cas, que de phonèmes cette mesure améliore son score ou si elle confirme un cas particulier de variabilité. Dans le contexte de voix journalistique cet indice est parmi ceux qui montrent le moins de variabilité inter locuteur, les meilleurs résultats étant autour de 10% pour les trois voyelles nasales, nous pouvons affirmer que dans ce genre d'enregistrements l'incidence est très limitée et ne donne pas lieu à des suggestions pour l'application future du HNR dans un contexte non contrôlé.

5.3 Durée et fréquence fondamentale

La prochaine mesure dont nous allons examiner les résultats est la durée, celle-ci n'apportant aucun effet significatif dans FABIOLE, nous nous concentrons donc uniquement sur les valeurs de l'autre corpus à notre disposition. Globalement la durée ne semble pas constituer un indice significatif de variabilité inter locuteur, 4 segments (/ʁ/, /f/, /v/, /z/) dépassent la valeur de 0.20 de la taille d'effet, la voyelle nasale /ã/ présente, aussi pour cette mesure, une taille d'effet significative, 65% dans ce cas, avec la même valeur pour les locuteurs et 50% pour les locutrices. La fricative /ʃ/ obtient aussi une valeur très élevée, 72% globale, cet effet ne se distribue pas de la même manière entre les deux sexes, nous remarquons une différence importante entre locuteurs 0.65 et locutrices 0.11.

La fréquence fondamentale étant une mesure pertinente uniquement pour les segments voisés, ils seront les seuls que nous allons considérer pour cet indice et montrer dans le Tableau 3 ci-dessus. Certaines valeurs ont été obtenues pour la f_0 aussi dans les segments non voisés, notamment /p/, /t/, /k/, /f/, /ʃ/, ceci est probablement dû à la segmentation de ces phonèmes parfois légèrement large lorsqu'ils se trouvaient proches de voyelles ou d'autres segments voisés.

f_0	**b**	**d**	**l**	**m**	**ʁ**	**v**	**z**
	83 (57 ; 70)	80 (54 ; 67)	60 (22 ; 79)	91 (77 ; 82)	29 (9 ; 67)	61 (27 ; 43)	72 (40 64)
	ã	**a**	**i**	**u**			
	84 (41 ; 79)	85 (65 ; 82)	84 (62 ; 72)	85 (64 ; 77)			
f_0 min	**b**	**d**	**l**	**m**	**ʁ**	**v**	**z**
	78 (51 ; 69)	70 (40 ; 67)	66 (31 ; 79)	92 (80 ; 82)	52 (26 ; 6)	59 (29 ; 43)	69 (40 ; 64)
	ã	**a**	**i**	**u**			
	86 (53 ; 90)	78 (51 ; 79)	82 (61 ; 67)	84 (63 ; 70)			
f_0 max	**b**	**d**	**l**	**m**	**ʁ**	**v**	**z**
	80 (60 ; 37)	72 (61 ; 26)	59 (27 ; 67)	89 (79 ; 57)	40 (20 ; 15)	63 (29 ; 76)	59 (33 ; 29)
	ã	**a**	**i**	**u**			
	76 (47 ; 60)	85 (65 ; 81)	83 (61 ; 70)	84 (61 ; 77)			

Tableau 3 Pourcentage de la taille d'effet selon le η^2 pour l'estimation de la f_0 - PATATRA

Cependant ces valeurs peuvent aussi indiquer un voisement non volontaire de la part du locuteur et pour cela nous laisserons les résultats correspondant à l'intérieur du tableau général présent en annexe. Les valeurs de f_0 dans notre premier corpus, mis à part /ʁ/ avec 0.29, montrent en général des effet égaux ou supérieurs à 60%, avec le segment /m/ qui a le plus haut à 91% suivi de /a/ et /u/ à 85% et les deux autres voyelles /i/ et /ã/ à 84%. Pour le corpus FABIOLE les résultats ne sont pas si élevés, nous retrouvons également /m/ qui recueille l'effet le plus haut avec 25%, suivent /l/ avec 24%, /v/ et la nasale notée comme /ɛ̃/ à 23%. Dans les deux cas les segments suivant sont les occlusives /b/ et /d/, nous

remarquons qu'en général les nasales, d'abord les consonnes puis les voyelles, maintiennent les scores de discrimination les plus élevés. En parole journalistique la fricative voisée joue un rôle plus important que dans l'exemple de parole spontanée alors que pour cette dernière ce sont les voyelles qui aident le plus dans l'analyse de la discrimination inter locuteur. Les résultats des voyelles pour le corpus PATATRA ne montrent pas de grandes différences entre ouvertes et fermées alors que pour l'autre base ce sont les voyelles ouvertes qui montrent un effet plus important, 21% pour /i/ et 20% pour /u/, même score que les deux autres nasales.

f_0	b	d	l	m	ʁ	v	z
	22	22	24	25	9	23	18
	ã	ɛ̃	ɔ̃	a	i	u	
	20	23	20	11	21	20	
f_0 min	b	d	l	m	ʁ	v	z
	21	22	22	26	8	23	18
	ã	ɛ̃	ɔ̃	a	i	u	
	21	23	22	11	21	20	
f_0 max	b	d	l	m	ʁ	v	z
	16	19	17	22	7	29	16
	ã	ɛ̃	ɔ̃	a	i	u	
	15	17	15	7	16	15	

Tableau 4 Pourcentage de la taille d'effet selon le η^2 pour l'estimation de la f_0 - FABIOLE

Nous avons effectué deux autres mesures de fréquence fondamentale que nous allons reporter ici, la première correspond à la fréquence fondamentale minimale. Les résultats restent assez proches de ceux que nous venons de décrire pour la fréquence fondamentale. Le segment avec l'effet le plus élevé dans le corpus de parole spontanée est /m/ avec 92% (90 pour les locuteurs et 57 pour les locutrices), suivi de la voyelle nasale avec 86% et de /u/ avec un η^2 de 84%, /v/ et /ʁ/ sont les seuls segments qui ne dépassent pas 60% d'effet sur le facteur locuteur pour cette indice. En ce qui concerne le corpus FABIOLE nous observons que /m/ a également le segment avec le score le plus élevé avec 26%, suivi de /v/ et /ɛ̃/ à

23%, et /ɔ̃/, /l/ et /d/ qui obtiennent la troisième valeur avec 22%. Les consonnes nasales restent aussi en terme de fréquence minimale les segments les plus discriminants, parmi ceux considérés, avec les voyelles nasales et orales qui viennent juste après suivies des occlusives. La latéral /l/ et la fricative uvulaire /ʁ/ se confirment globalement comme les variables les moins discriminantes pour les mesures de f_0 même si les effets qu'elles montrent ici, 52% pour /ʁ/ et 66% pour /l/, représentent les valeurs les plus élevées parmi les trois mesures prises en compte. Cependant nous remarquons que le segment latéral montre pour les locuteurs des scores très élèves : 79% dans f_0 et f_0 min et 67% dans f_0 max.

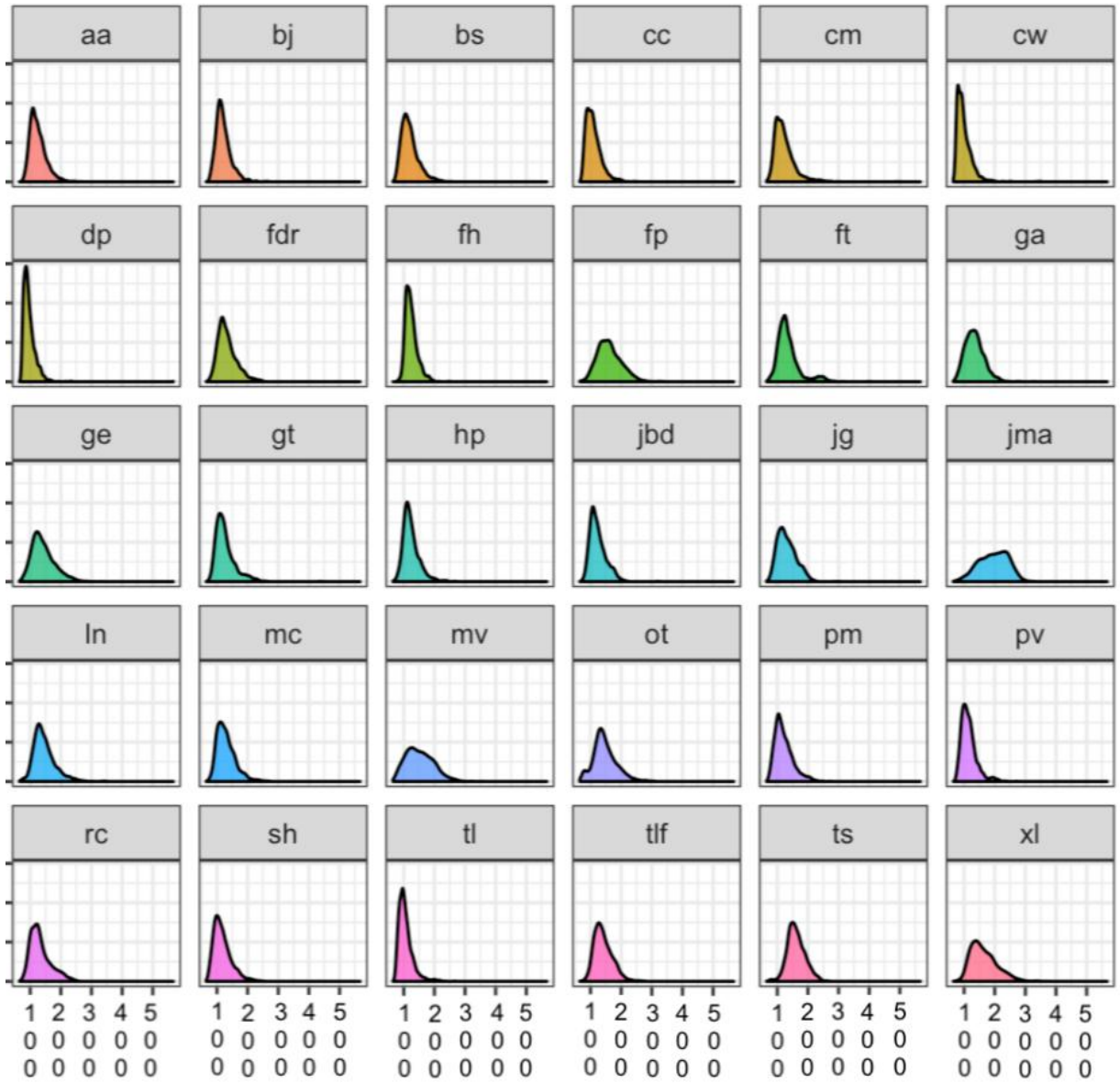

Fig. 6 valeurs de f_0-min (Hz) pour /m/ par locuteur dans le corpus FABIOLE sous forme de densité spectrale

La dernière mesure sur la fréquences fondamentale est la fréquence maximale. Cette fois encore pour les deux corpus c'est le segment /m/ qui obtient le score de η^2 le plus élevé, 89% et 22%. Pour la base FABIOLE seule la fricative /v/ montre un effet supérieur à 0.20, uniquement l'occlusive /d/ s'y approche avec 19% alors que les autres phonèmes restent en dessous de cette valeur. Pour le PATATRA les autres valeurs pour /a/ correspondent à 85%, 84% pour /u/, 83% pour /i/ et 80% pour /b/, 76% pour la nasale qui réalise le score minimale des voyelles pour cette mesure, les deux autres segments qui montrent un score supérieur à 60% sont /d/ avec 72% et /v/ avec 63%. Les valeurs de la fréquence fondamentale correspondent à des résultats significatifs dans les deux corpus, ceci montre l'efficacité de cette mesure indépendamment de l'annotation ou du contexte.

Dans la Figure 6, seul le segment /m/ pour le corpus FABIOLE est représenté, toujours à l'aide de la fonction de densité spectrale, pour montrer en quelle mesure ces résultats sont représentatifs de la variabilité inter locuteur puisqu'il s'agit de l'indice qui obtient les meilleurs résultats. Nous observons que les locuteur présentent des courbes très différentes : chez certains les pics sont plus prononcés, ce qui indique des observations plus nombreuses sur ces valeurs de fréquences. Les motifs des densités placés vers la gauche des figures représentent des valeurs minimales de fréquence fondamentale autour des 100 et 200 Hz, uniquement un locuteur, jma, sixième densité de la troisième ligne, présente des valeurs légèrement supérieures aux autres individus.

Dans les deux corpus nous observons une représentation plus ou moins particulière propre à chaque locuteur pour les valeurs de fréquence fondamentale, cependant les résultats obtenus ne semblent pas encore pouvoir expliquer à eux seuls la variation. Des mesures plus précises pourraient être appliquées par la suite de façon à affiner les calculs et réduire les espaces de valeurs communs aux locuteurs qui nous portent à affirmer la bonne réussite de la f0 mais pas encore son efficacité.

5.4 Mesures de rythme

Les résultats de η^2 que nous obtenons pour les mesures de rythme dans le corpus FABIOLE tournent autour de 10% d'effet avec les meilleurs valeurs correspondant à 0.16 et

0.17 pour les mesures brutes globale, rPVI et vocalique, VrPVI. Pour l'autre corpus analysé, l'indice normalisé consonantique a un effet mineur avec 26% alors que toutes les autres mesures montrent un effet sur le facteur locuteur supérieur à 50%, avec les deux meilleures valeurs, 0.60, qui correspondent aux mêmes indices que pour le corpus FABIOLE.

Comme nous l'avions anticipé lors de la présentation des corpus, les effets reportés par certaines mesures pouvaient se différencier considérablement entre les deux bases, c'est le cas ici des mesures de rythme et ce n'est pas étonnant : dans le corpus PATATRA les tests de PVI ont été effectués sur des extraits dans lesquels les locuteurs étaient amenés à lire un texte à la vitesse de leur choix, ceci nous fournit des enregistrements de parole plus spontanée que ceux provenant de la base FABIOLE où la parole journalistique qui est enregistrée montre une similarité majeure dans le rythme et la prosodie des locuteurs.

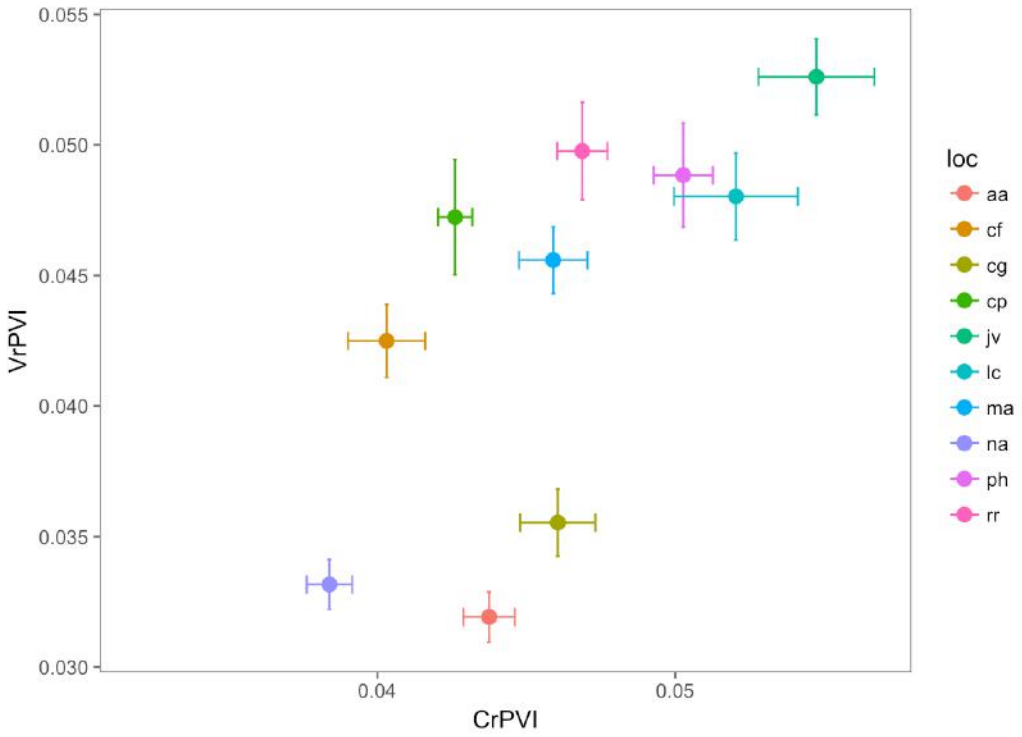

Fig. 7 combinaison CrPVI-VrPVI dans les corpus PATATRA, chaque point correspond à un locuteur

Nous avons aussi vu précédemment que les mesures de PVI sont souvent utilisées de manière couplée dans la littérature, CrPVI-VnPVI résultant comme la combinaison la plus efficace selon (Gharsellaoui et al., 2018 ; Toussaint, 2013) et d'après (Nolan et Asu, 2009) l'utilisation de l'indice brut est plus pertinent au niveau consonantique et l'indice normalisé

pour les voyelles. Cependant les scores que nous obtenons sont en contraste avec cette dernière affirmation : l'indice brut est plus efficace que celui normalisé dans les deux contextes et aussi globalement. La combinaison CrPVI-VrPVI se trouve être plus pertinente pour notre analyse, les Figures 8 et 9 exposent les résultats pour les deux corpus. Nous observons que les locuteurs du corpus contrôlé, pour lesquels les extraits proviennent d'une tâche de production de parole spontanée en chambre sourde, s'éloignent visiblement les uns des autres et dans un seul cas deux locuteurs se rapprochent.

Le milieu contrôlé des enregistrement favorise la qualité de ceux-ci mais nous observons que les locuteurs non contrôlés sont aussi distribués de manière homogène dans l'espace. Plusieurs représentations de locuteurs dans le cas du corpus FABIOLE se chevauchent mais la majorité arrive à s'éloigner de manière évidente. Le rythme comme la fréquence fondamentale fait parti des indices particuliers au locuteur et aux stratégies de production de la parole selon ce que nous avons pu observer, ces indices se trouvent être efficaces dans les deux contextes d'enregistrement.

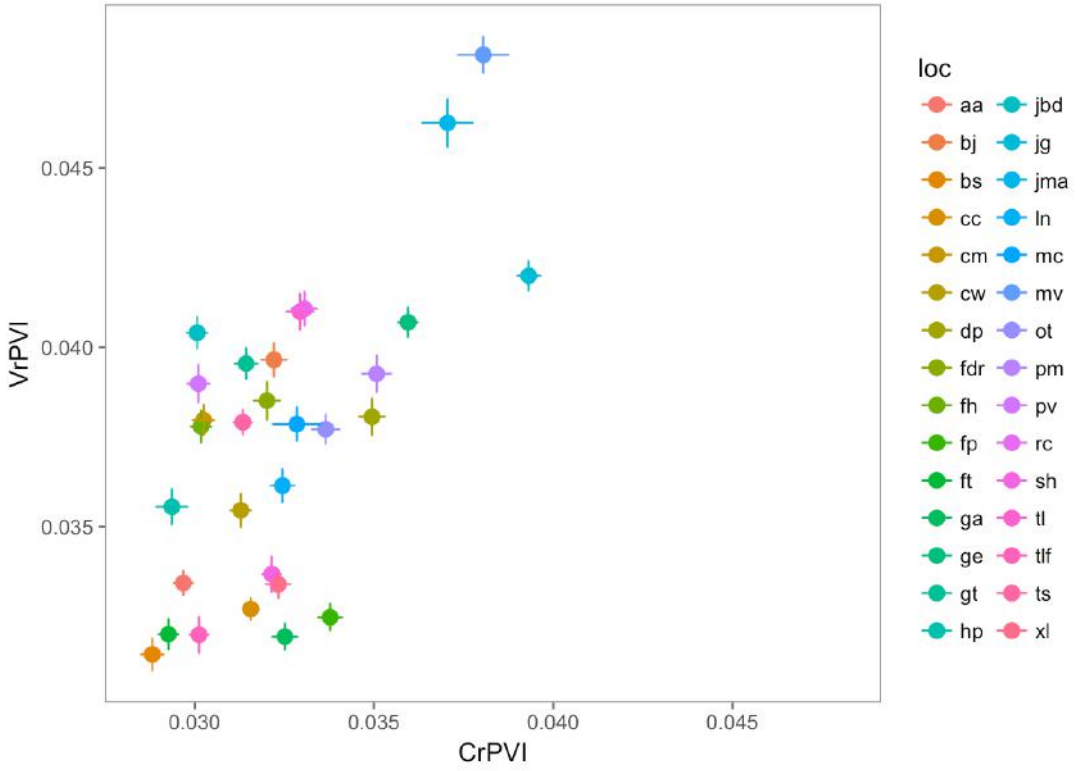

Fig. 8 combinaison CrPVI-VrPVI dans les corpus FABIOLE, chaque point correspond à un locuteur

Le contexte de voix journalistique aurait pu poser des problèmes car nous aurions pu retrouver des voix et des rythme d'énonciation très similaires, dû par exemple à des facteurs provenant de contraintes temporelles à l'intérieur des émissions mais les résultats ont démenti cette hypothèse et au contraire ils se sont montrés plutôt prometteurs. D'autres mesures de rythme pourraient apporter plus d'informations sur le locuteur et des calculs plus précis peuvent être mis en place pour exploiter ce facteur.

5.5 MFCC

Les derniers résultats correspondent aux coefficients cepstraux, comme mentionné précédemment cet ensemble se compose dans notre cas de 13 coefficients. Ce choix est véhiculé par la littérature citée où cette mesure a été retrouvée, notamment (Kamruzzaman et al., 2010 ; Spinu et Lilley, 2016) qui utilisent des ensembles de 13 coefficients, chez (Davis et Mermelstein, 1980) nous retrouvons l'utilisations de deux sous-ensembles de 6 et 8 coefficients à partir d'un vecteur global de 14. Nous avons également l'exemple de (Sahidullah et Saha, 2013) qui utilisent 12 coefficients et (Ahmad et al., 2015) qui prennent des vecteurs variables entre 14 et 19 coefficients alors que (Kahn, 2011) utilise le vecteur le plus large de la littérature que nous avons étudié avec 20 coefficients cepstraux. Ces indices sont habituellement utilisés en reconnaissance du locuteur (voir état de l'art) pour modéliser les locuteurs sous forme de vecteur, n'étant pas dans le cadre d'une analyse à l'aide d'un système de reconnaissance et comme anticipé en début de chapitre nous analyserons leur effet séparément, les prenant de manière statique comme pour les autres indices acoustiques.

Pour le corpus FABIOLE, nous obtenons globalement des résultats moyens pour les nasales et le /a/, aucun segment arrive à dépasser le seuil du 60% d'êta carré. Les occlusives représentent les phonèmes pour lesquels les MFCC sont un indice de variabilité inter locuteur peu pertinent. Le meilleur score est obtenu par les occlusives bilabiales /p/ /b/ avec le deuxième coefficient, le η^2 s'élève à 0.20 et 0.22. Les valeurs les plus représentatives correspondent aux trois voyelles nasales présentes dans ce corpus /ã/, 41%, /ɔ̃/, 38%, les deux dans le coefficient 7, /ɛ̃/ obtient le score de 35% dans le premier coefficient et ceci est le résultat majeur de ce segment. Le triplet de voyelles nasales est

toujours présent avec des valeurs supérieures à 20% de la taille d'effet, sauf dans le coefficient 5 où seul /ã/ /õ/ /m/, la consonne nasale /m/ dans le coefficient 5 montre son majeur score avec 33% et qui registre uniquement dans les coefficients 10 et 13 des valeurs inférieures à 20%. La voyelle ouverte /a/ présente aussi lors de 8 coefficients des valeurs d'êta carré supérieures à 0.20. Avec le treizième coefficient nous observons les tailles d'effet les moins importantes, aucun segment, y compris les nasales, ne dépassant le 20% de η^2. Deux autres segments montrent enfin des valeurs de taille d'effet supérieures à 20% dans des cas particuliers, pour la fricative /ʃ/ il s'agit du deuxième coefficient et pour la latérale /l/ nous l'observons dans les coefficients 2, 4 et 5.

Phonème	FABIOLE		PATATRA	
	Coefficient	η²	Coefficient	η²
p	MFCC2	20	MFCC2	10 (6 ; 8)
t	MFCC2	15	MFCC3	11 (9 ; 1)
k	MFCC2	12	MFCC3	9 (9 ; 6)
b	MFCC2	22	MFCC7	19 (18 ; 23)
d	MFCC2	21	MFCC5	41 (40 ; 23)
l	MFCC2	15	MFCC6	39 (6 ; 24)
m	MFCC7	33	MFCC6	68 (65 ; 55)
ʁ	MFCC3	12	MFCC8	27 (17 ; 29)
f	MFCC3	14	MFCC1	40 (42 ; 31)
ʃ	MFCC3	25	MFCC5	45 (39 ; 52)
v	MFCC2	14	MFCC12	34 (18 ; 6)
z	MFCC5	12	MFCC1	47 (44 ; 6)
ã	MFCC7	41	MFCC12	81 (81 ; 11)
ɛ̃	MFCC1	35	/	/
ɔ̃	MFCC7	38	/	/
a	MFCC1	29	MFCC9	56 (60 ; 24)
i	MFCC5	24	MFCC12	68 (28 ; 14)
u	MFCC3	16	MFCC9	55 (46 ; 48)

Tableau 5 Récapitulatif des meilleures tailles d'effet selon le η^2 pour les MFCC dans les deux corpus

Le corpus PATATRA présente des valeurs globales légèrement plus élevées, ceci n'est pas très étonnant selon les considérations faites tout au long de notre analyse, les différences en terme de nombre de locuteurs et type d'enregistrement jouant un rôle important sur l'efficacité des indices considérés. Une seule voyelle nasale /ã/ était produite par les 10 locuteurs et elle est le segment qui montre les scores les plus élevés, arrivant dans 6 coefficients à réaliser un êta carré supérieur à 0.60. La valeur la plus élevée correspond au douzième coefficient cepstral, avec 81% elle est le segment avec la valeur la plus importante pour l'indice MFCC dans ce corpus.

Nous remarquons aussi une différence importante entre les résultats obtenus pour les locutrices et ceux pour les locuteurs, par exemple avec 81% réalisé par /ã/ dans le coefficient 12 partagé par les femmes uniquement, chez les hommes ce segment réalise un score de 11%. Les résultats de taille d'effet obtenus par les segments considérés montrent des valeurs plus élevées pour les locutrices lors de l'analyse des coefficients cepstraux pour les segments nasals, voyelle /ã/ et consonne /m/. La voyelle ouverte /a/ montre un résultat plus significatif pour les locuteurs, tous les autres segments ne montrent pas de déséquilibre vers un sexe plutôt que vers l'autre, ayant globalement des valeurs la plupart du temps constantes entre les deux sous-ensembles.

D'autres segments dépassent le seuil de 0.60 en terme de η^2, notamment la consonne nasale /m/ et la voyelle fermée /i/ montrent une taille d'effet de l'ordre de 68% respectivement dans les coefficients 6 et 12. L'autre voyelle fermée /u/ avec la voyelle ouverte /a/ peuvent être considérées comme les meilleurs segments ayant une taille d'effet moyenne, ils ne dépassent pas le seuil du 0.60 de η^2, mais leurs valeurs oscillent toujours entre 35% et 55%. Les fricatives montrent aussi des résultats moyens autour de 40%, avec le segment /ʒ/ qui réalise le meilleur score du coefficient 13 avec 47%. Enfin nous remarquons comment les occlusives non voisées sont les segments les moins significatifs en ce qui concerne les coefficients cepstraux, leur taille d'effet ne dépassant jamais 10%. Cependant, l'occlusive voisée /d/ arrive à réaliser un score de 41% dans le coefficient 5. Le tableau 5 résume les meilleurs scores de taille d'effet obtenus par les coefficients cepstraux dans les deux corpus, dans la colonne de PATATRA seront inscrites les valeurs globales ainsi que entre parenthèses celles correspondant aux locutrices et locuteurs.

5.6 Synthèse

Nous allons maintenant procéder à une récapitulation de ce que les résultats obtenus ont pu nous dire et à la comparaison entre les valeurs de êta carré déjà présentes dans un travail antérieur similaire (Kahn, 2011), utilisant le corpus BREF constitué d'enregistrements téléphoniques. Nous nous limiterons dans ce paragraphe aux valeurs les plus représentatives, des tableaux détaillés seront présents en annexe. Nous ferons aussi une distinction nette entre les locutrices et les locuteurs puisque le corpus analysé BREF était aussi divisé par sexe, nous tenons aussi à rappeler que nos locutrices ne constituent qu'un groupe de 5 par conséquent ces résultats devront être considérés limités.

Dans les deux corpus à notre disposition nous remarquons une grande efficacité des segments nasals indépendamment du sexe des locuteurs : /ã/ obtient le score le plus élevé, 90% et 41% pour les hommes respectivement dans PATATRA et dans FABIOLE qui partage la valeur avec BREF ; 51% dans BREF contre 81% pour notre groupe de locutrices. Dans chaque corpus l'indice utilisé est différent, f_0-min dans PATATRA hommes, le coefficient cepstral 7 dans FABIOLE et le coefficient 12 dans PATATRA femmes, seul le corpus BREF maintient constant l'indice du centre de gravité spectral qui s'est montré dans ce travail comme le plus influencé par le facteur locuteur. Les deux autres voyelles nasales /ɛ̃/ et /ɔ̃/, absentes du corpus PATATRA, montrent respectivement le deuxième, 38% sur le septième coefficient cepstral, et troisième, 37%, meilleurs résultats, cette fois sur le centre de gravité à puissance 1.

Inversement dans BREF nous observons que l'effet du locuteur sur le centre de gravité de /ɔ̃/ est supérieur avec 53% contre 45% pour /ɛ̃/. Le dernier segment nasal, la consonne bilabiale /m/ obtient le meilleur résultat pour les segments consonantiques, à travers le septième coefficient cepstral dans FABIOLE la valeur de η^2 est de 24%, le centre de gravité dans BREF résulte en 39% pour les homme et 23% pour les femmes alors que dans PATATRA nous obtenons respectivement 84% et 80% grâce à la mesure de f_0-min.

Parmi les segments ne présentant pas le trait de nasalité nous retrouvons notamment /l/ comme la consonne obtenant les valeurs les plus élevées : pour les hommes 22% dans BREF, 24% dans FABIOLE et 79% dans PATATRA, à travers le centre de gravité pour le premier corpus et la fréquence fondamentale pour les deux autres ; le groupe de locutrices

obtient une valeur de 17% dans BREF et nous obtenons 31%. Parmi les voyelles cardinales, celle qui montre la valeur de η^2 la plus élevée dans tous les corpus est /a/ avec 45% selon les variations du quatrième formant chez les hommes dans BREF, 30% dans FABIOLE à travers la mesure du centre de gravité spectrale à puissance 1, le même indice donne 51% pour BREF chez les locutrices. Dans PATATRA la f_0 montre 82% de taille d'effet pour les hommes et 65% pour les femmes. Il est intéressant de voir comment, lorsque nous sommes en milieu contrôlé cette voyelle se trouve plus efficace pour expliquer la variabilité inter locuteurs de sexe masculin alors que dans un milieu non contrôlé la tendance est opposée ayant un effet légèrement majeur sur les locutrices. Les deux autres voyelles gardent des valeurs qui oscillent entre 25 et 30% dans les corpus en milieu non contrôlé pour les deux sexes, les valeurs dans PATATRA sont similaires à celle de /a/ pour les femmes et légèrement inférieures pour les hommes.

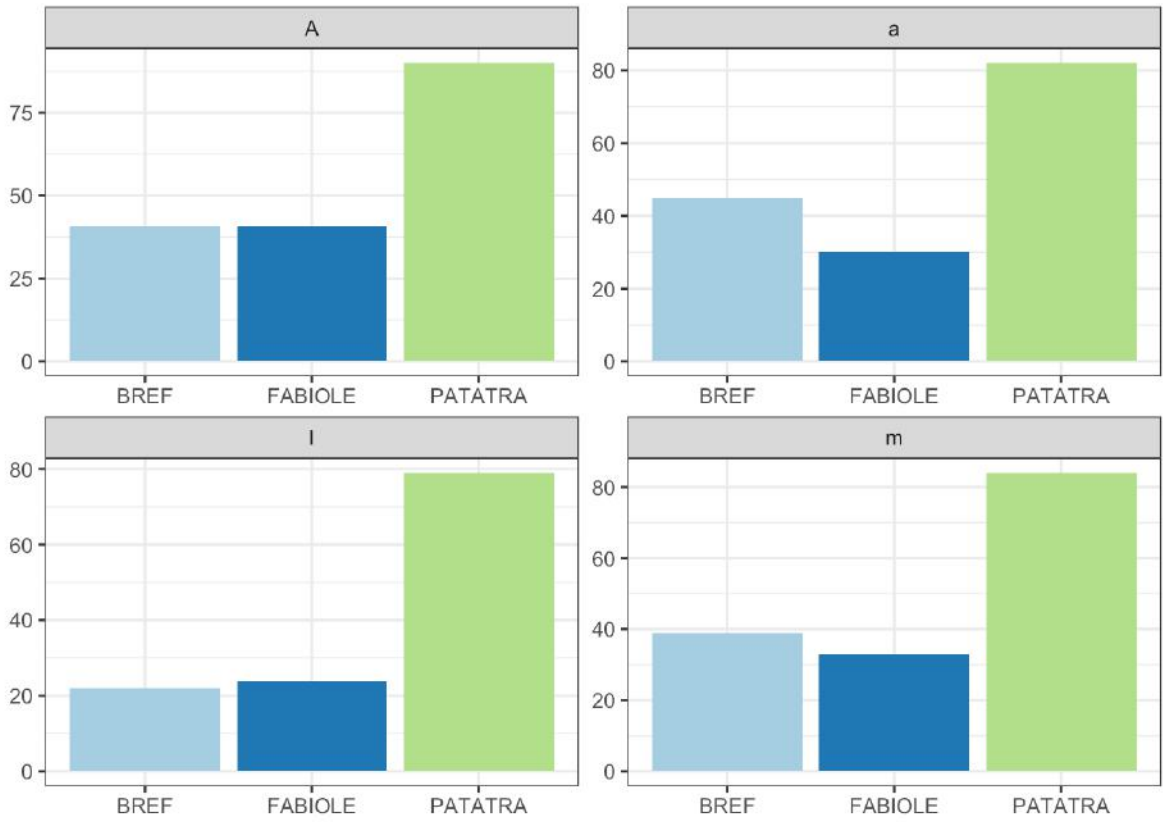

Fig. 9 comparaison des η^2 entre les corpus BREF, FABIOLE, PATATRA-hommes pour /ã/ /m/ /a/ /l/

Dans la Figure 10 nous montrons les différences d'effet dans les trois corpus d'hommes pour les meilleurs quatre consonnes et voyelles avec un segment oral et un nasal pour

chaque catégorie. Nous pouvons observer que les résultats obtenus pour les enregistrements non contrôlés sont constants ou légèrement inférieurs à ceux de (Kahn, 2011), sur les segments déjà observés comme étant les meilleurs notre apport est très réduit puisque avec des nouveaux paramètres obtenons les mêmes scores. La recherche d'indices pour expliquer la variabilité inter locuteur pour ces phonèmes doit se complexifier pour obtenir des mesures plus précises, capables d'aller au-delà de ces résultats. Le cas des enregistrements en milieu contrôlé nous montre des éléments largement plus représentatifs, mais encore une fois nous devons considérer les limitations en terme de nombre de locuteurs que ce corpus nous offre, ceci peut nous pousser par la suite à poursuivre ce genre d'analyse sur un corpus présentant les mêmes caractéristiques mais plus large et varié pour identifier des éléments qui peuvent être appliqués en suite au travail en milieu non contrôlé.

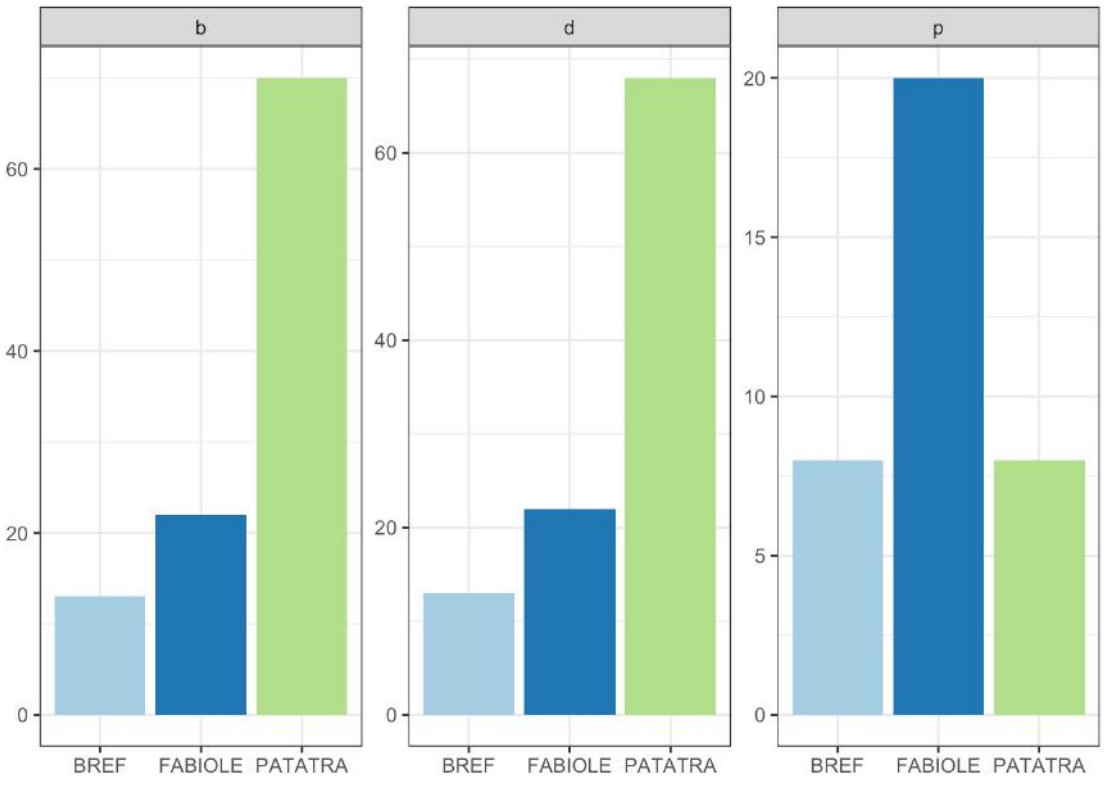

Fig. 10 comparaison des η^2 entre les corpus BREF, FABIOLE et PATATRA-hommes pour /p/ /b/ /d/

D'autres segments montrent une amélioration de l'effet du locuteur entre le corpus BREF et les nôtres comme montré dans la Figure 11, notamment les occlusives /p/ /b/ /d/, pour les voisées ce sont les valeurs de f_0 et f_0-min les plus porteuses d'informations sur le locuteur

alors que pour la bilabiale non voisée c'est le deuxième coefficient cepstral. Pour les fricatives nous n'observons pas de différences remarquables entre les valeurs obtenues par nos indices et celles déjà discutées. Les phonèmes voisés arrivent à presque doubler les résultats de comparaison avec notre corpus de voix journalistique et /p/ obtient un score trois fois plus important.

En milieu contrôlé ce segment ne montre pas la même efficacité que les deux autres. Les occlusives représentent le résultat le plus surprenant de notre étude, l'amélioration des scores de ces phonèmes est évidente et cela permet de confirmer ce qui avait été dit au début de notre étude lorsque nous avions établi l'ensemble des segments considérés comme porteurs d'informations, à savoir voyelles nasales, consonnes nasales, voyelles orales, fricatives, plosives et approximantes. La présence des plosives se confirme et ces phonèmes se trouvent avoir dans notre cas des scores plus élevés que les fricatives.

Ainsi les indices porteurs d'informations particulières sur les locuteurs qui pourront être approfondi pour des futurs travaux d'analyses se différencient selon les phonèmes et nous aident à comprendre dans quelles mesures chaque production langagière est caractérisée de manière propre.

Cette étude représente une analyse plus approfondie de paramètres autres que ceux présents dans la littérature de la reconnaissance du locuteur. Nous avons cherché, à travers un ensemble large de mesure acoustique et de coefficients cepstraux, à identifier quels éléments sont les plus pertinents au niveau phonémique, ou dans l'ensemble de l'énoncé pour les mesures de rythme. Globalement nous retrouvons des résultats cohérent avec ce qui a été observé précédemment, les moments spectraux et le HNR apportent des informations de qualité inférieure par rapport à la fréquence fondamentale et au rythme. Les premiers révèlent des caractéristiques statiques dans le signal sonore alors que les autres mettent en évidence des caractéristiques propres aux locuteurs de manière dynamique pendant le processus d'énonciation. Les informations statiques et dynamiques caractérisent l'identité des locuteurs et peuvent être d'avantage exploitées.

6 Résultats et discussion

Dans ce chapitre notre attention ne sera plus portée sur la variabilité interphonémique, que les indices pris en examen arrivent à expliquer, mais sur le rôle de ces indices dans la caractérisation des locuteurs. Nous avons en partie déjà pu voir cela dans le chapitre précédent, en particulier lors de la description des résultats provenant des mesures de rythme, les locuteurs étaient représentés dans un espace et leur position décrivait en quelle mesure les indices de PVI, les mesures brutes en contexte vocalique et consonantique dans le cas de la Figure 8, pouvaient être représentatifs d'un locuteur. D'autres représentations des locuteurs dans des espaces de paramètres ont été mis en place et feront partie de la discussion qui va suivre. Comme nous l'avons anticipé dans le chapitre 5 nous nous servirons de techniques telles que l'Analyse en Composantes Principales et les arbres de décision pour interpréter les données à notre disposition.

6.1 PATATRA

Le premier ensemble de données que nous allons aborder est le corpus PATATRA : composé de 10 locuteurs (6 femmes et 4 hommes) et d'enregistrements effectués en milieu contrôlé (une liste de mots lue en chambre sourde avec microphone casque). La première représentation que nous fournirons est celle obtenue à travers l'Analyse en Composantes Principales. Nous avons procéder à l'application de cette méthode en deux étapes : la première transposition des données dans l'espace de l'ACP a été effectuée avec l'ensemble brut des valeurs contenues dans notre tableau alors que dans un deuxième temps nous avons effectué une normalisation pour observer l'impact que cela pouvait avoir sur les résultats.

Le nuage des individus en Figure 11 montre la représentation des données brutes, la variabilité du corpus n'est expliquée qu'à 53.54%, valeur combinée des deux Composantes Principales que nous retrouvons sur les deux axes. La distribution des individus ne se fait pas de manière homogène : les valeurs se distribuant tout au long de l'espace nous n'avons pas une représentation simplifiée et mieux interprétable, ce qui est pourtant l'objectif d'une ACP. Cependant la représentation du corpus est fidèle puisque nous observons en moyenne

des valeurs très proches et peu distinctives ainsi que des valeurs qui sortent complètement du corps du nuage pour désigner plus particulièrement certains locuteurs. Ceci explique le pourcentage de variabilité à plus de 50% que cette représentation arrive à nous fournir. Ce pourcentage n'est pas le seul élément qui nous aide à décrire le corpus lors de l'Analyse en Composantes Principales, nous obtenons aussi le résultat de la contribution de chaque variable pour l'espace considéré. Ce résultat nous fournit deux indicateurs sur les indices considérés nous expliquant en quelle mesure chaque indice influence la distribution de la population et par conséquent lequel est plus discriminant dans cette représentation. Un tableau récapitulatif des contributions pour chaque ACP est présent en Annexe. Pour ce premier ensemble de données les deux variables qui contribuent le plus à la description de la variabilité sont les valeurs maximales et minimes de la fréquence fondamentale ayant les deux une contribution autour du 18%. Les autres indices se partagent le pourcentage restant, avec l'écart type du HNR qui est le seul à ne pas dépasser le 0%.

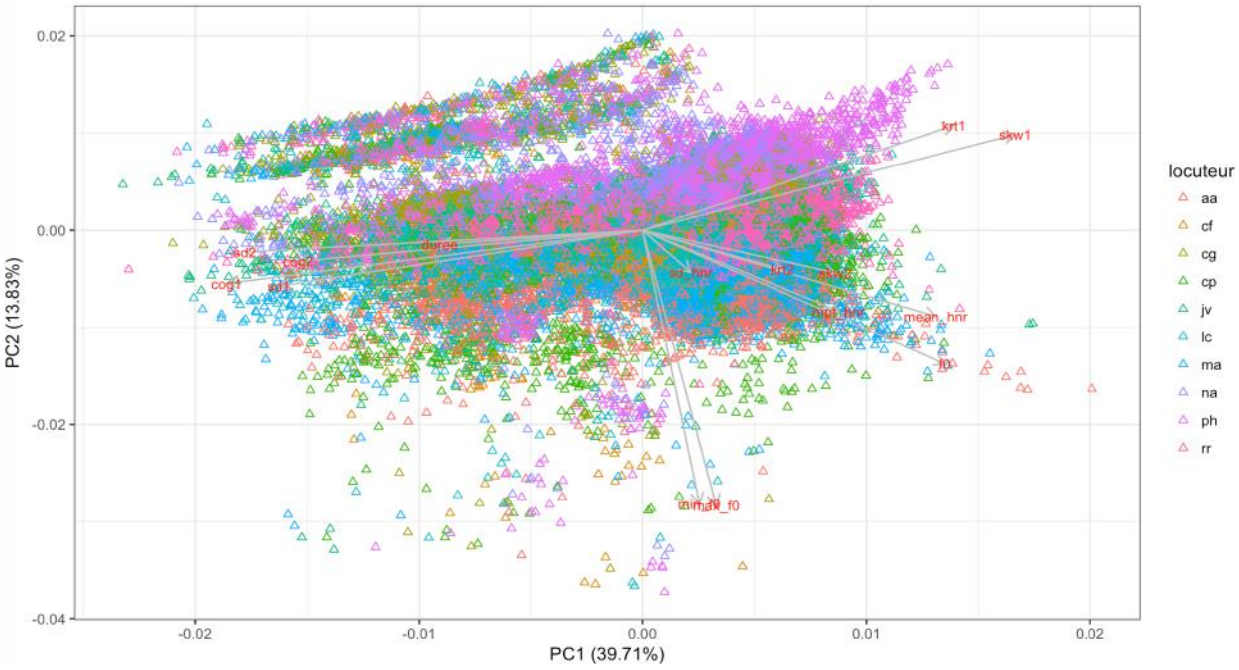

Fig. 11 représentation par Composantes Principales 1 et 2 pour du corpus PATATRA : données brutes ; nuage des variables et des individus simultanément

L'autre élément remarquable concerne la projection des variables dans cet espace. Nous rappelons que la direction des vecteurs des variables n'influence pas la distribution de la population mais nous donne des indications sur les valeurs de ces indices. Les valeurs qui suivent la trajectoire de ces vecteurs seront plus conséquentes pour les locuteurs

représentés : nous observons une population réduite au-delà de la projection des vecteurs de f0, cela nous indique que ces locuteurs-là seront caractérisés par des indices de fréquence fondamentale plus élevée que la moyenne générale. Nous pouvons aussi faire des considérations sur la corrélation entre ces variables : les vecteurs allant dans des directions opposées indiquent une relation inversement proportionnelle de ces indices. Nous observons ce comportement pour les deux premiers moments spectraux, à savoir centre de gravité et écart type, qui s'opposent aux deux autres, skewness et kurtosis. Les locuteurs présentant des valeurs plus élevées pour les premiers MS auront des résultats nettement inférieurs dans les deux autres. De la même manière nous observons une opposition entre la durée et les indices du rapport Harmoniques sur Bruit.

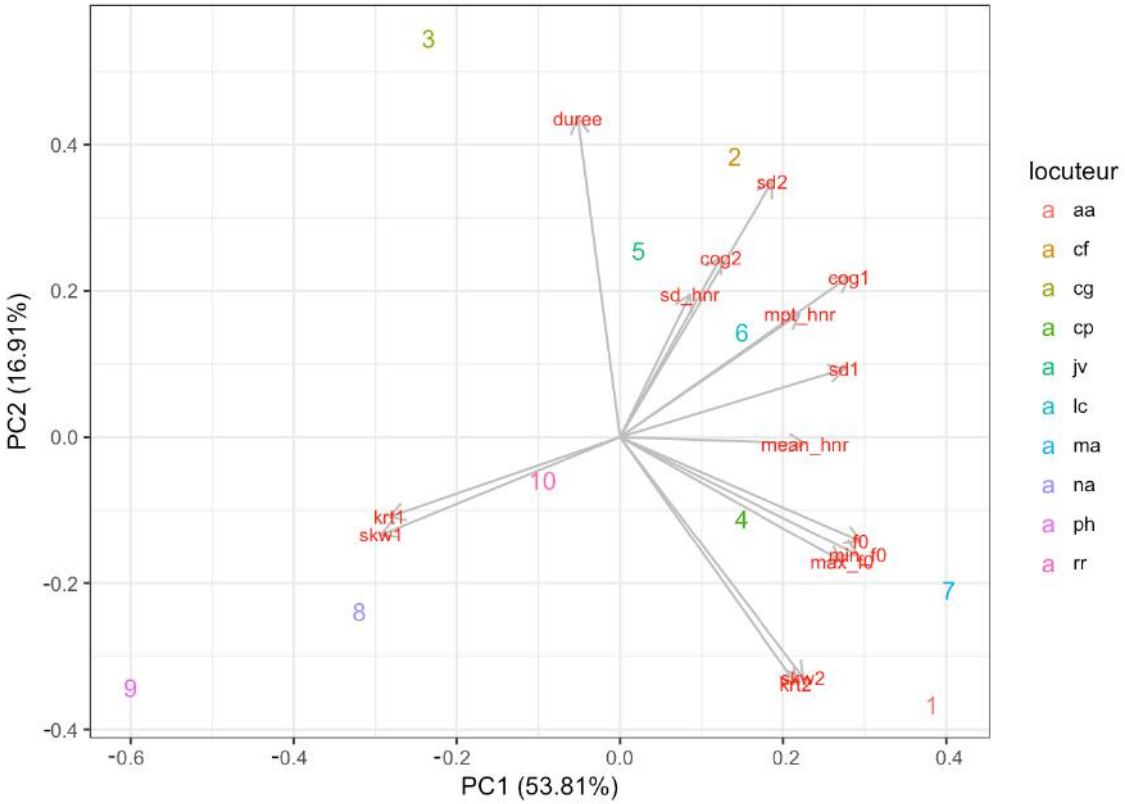

Fig. 12 représentation par Composantes Principales 1 et 2 pour du corpus PATATRA : données normalisées ; nuage des variables et des individus simultanément

La deuxième description de l'Analyse en Composantes Principales sera faite à partir de données normalisées : la moyenne pour les valeurs de chaque locuteur est calculée. À travers les deux premières composantes qui se trouvent être celles qui regroupent la majorité de la variabilité, nous obtenons une représentation cumulée de 70.72% de la

variabilité inter locuteur : la première composante explique 53.81% alors que la deuxième 16.91%. Les variables dans ce cas se partagent équitablement le pourcentage de contribution avec seuls les mesures de HNR, le centre de gravité spectral en puissance 2, le kurtosis et l'écart type en puissance 1 qui ont des taux inférieurs. Dans la Figure 12 nous avons : sur l'axe des abscisses la première Composante Principale (PC1) et sur l'axe des ordonnées la deuxième (PC2) ; les locuteurs sont représentés par les différents chiffres distribués de manière homogène dans l'espace ; les variables sont aussi projetées, en forme de vecteur, pour rendre compte de leur distribution par rapport au nouveau plan de représentation. Certains éléments intéressants ressortent de cette analyse.

Les locuteur 3, 8, 9 et 10 sont les seuls à avoir des valeurs négatives sur l'axe des abscisses alors que tous les autres possèdent des valeurs positives. Ce résultat nous fournit déjà une indication importante sur la distribution des individus puisque ces quatre locuteurs représentent les hommes de notre corpus. La Première Composante arrive à diviser de manière homogène le premier sous-ensemble sans avoir d'indication puisque les données d'entrée fournies à l'ACP n'étaient que les valeurs des différentes variables. Cette observation reste limitée par le nombre de locuteur à notre disposition et elle ne pourra pas être confirmée par le deuxième corpus contenant uniquement des locuteurs de sexe masculin, elle pourrait donc faire l'objet d'analyses ultérieures. La PC1 est la seule capable de nous donner l'information sur le sexe du locuteur après l'étude des autres espaces de représentation ayant comme axes les Composantes successives.

Comme pour l'exemple des données brutes, une autre indication que nous pouvons déduire à l'aide de cette représentation est la corrélation existante entre les différentes variables considérées. La direction des vecteurs représentant les variables dans l'espace nous indique principalement le degrés de corrélation entre ces éléments : nous comprenons que les deux centres de gravité spectrale et les écarts type se trouvent à l'opposé des coefficients d'aplatissement et d'asymétrie en puissance 1. Si nous rapportons cette considération aux locuteurs nous aurons pour un locuteur présentant des valeurs élevées dans un des premiers indices cités des valeurs moins conséquentes dans les deux autres. De la même façon nous observons le comportement de la durée opposée aux indices de f_0 et aux coefficients spectraux d'asymétrie et aplatissement en puissance 2. Comme cela a été dit dans le chapitre 5, les variables présentes dans notre corpus de départ n'influencent

pas la position des individus dans l'espace lors de l'Analyse en Composantes Principales mais elles sont projetées selon le même calcul, pour cela les positions des unes et des autres nous fournissent une indication sur la combinaison de ces éléments dans la représentation des locuteurs lors de l'ACP. Les locuteurs qui se placent au-delà de la projection vectorielle d'une variable seront interprétés comme ayant des valeurs plus importantes pour ces indices, cela nous porte à penser que ces individus seront plus influencés par ces variables lors de la classification. Par exemple, les locutrices 1 et 7 se trouvant au-delà des indices de f_0 nous pouvons nous attendre à ce que ces indices, lors de la prise de décision, aient un poids important dans la classification de ces deux individus. De la même manière cela devrait se produire pour l'indice de durée pour le locuteur 3. Nous confronterons ces hypothèses avec les résultats des arbres de décision pour essayer de comprendre en quelle mesure ces deux méthodes diffèrent dans la représentation de l'ensemble des individus.

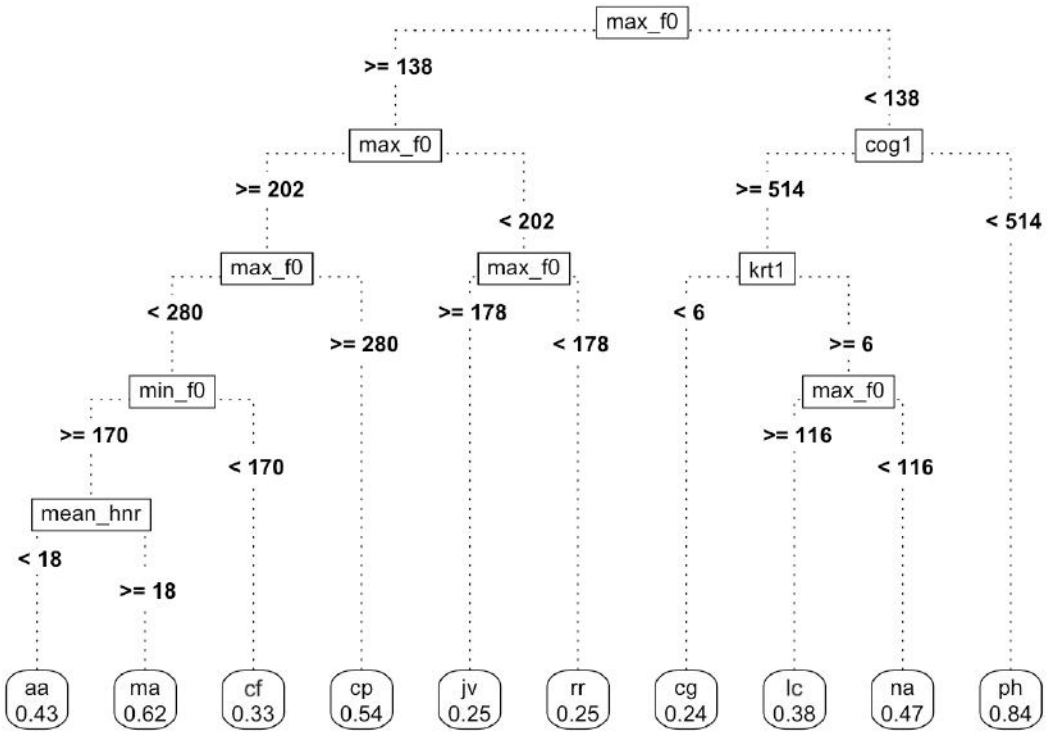

Fig. 13 arbre de classification pour le corpus PATATRA, données quantitatives

Nous confirmons grâce à l'arbre dans la Figure 13 une partie des hypothèses que nous venons de nommer. La variable de départ de l'arbre étant la valeur maximale de la fréquence fondamentale nous retrouvons tout de suite que les locutrices 1 et 7, aa et ma, ayant les valeurs les plus élevées de f_0 selon l'ACP, se placent à l'extrême gauche de l'arbre après trois branches qui divisent les valeurs de fréquence fondamentale, se distinguant entre elles grâce à l'indice de HNR. L'arbre de classification ne reporte pas d'indication sur la durée, le locuteur 3 se distingue à travers deux mesures spectrales. La différence la plus saillante entre l'ACP et l'arbre de classification est la division du corpus par sexe : si elle était très précise dans la première, cette fois nous n'avons pas d'indicateur marquant cette différence, ne s'agissant que de données quantitatives. Nous observons lors de la prise de décision que dans l'arbre les indices de fréquence fondamentale sont ceux qui ont le plus d'influence dans la discrimination des locuteurs et ceci de manière très articulée : 6 nœuds sur 9 reportent des valeurs sur cet indice ; 6 extrémités sur 10 sont également issues d'une discrimination provenant de valeurs de f_0.

La présence importante de cette variable lorsque les informations extraites du signal sonore sont gardées intactes porte à affirmer que l'influence de la fréquence fondamentale joue un rôle très important dans la discrimination des locuteurs. Cependant cet indice n'arrive pas à décrire toute la variabilité de la population pour cela nous retrouvons donc des indices complémentaires qui apportent des informations supplémentaires. Lorsque nous excluons les indices principaux de la classification, les variables jouant un rôle complémentaire deviennent principales mais n'arrivent pas à transcrire la totalité de la variabilité du corpus. Ceci rend les variables complémentaires même si plus nombreuses bien plus faibles en terme de représentation de la population.

Une autre moyen que nous avons utilisé pour représenter des locuteurs à travers les valeurs des indices extraits dans le signal sonore est celui des diagrammes en radar. Cette représentation graphique conserve toutes les informations propres aux valeurs des locuteurs qui peuvent être présentées sur un espace pluridimensionnel, en créant conséquemment des graphes uniques pour chaque locuteur si nous avons à faire à des données ayant une variation de base très importante. Dans notre cas nous assistons globalement à des représentations peu explicatives car la variabilité des indices ne permet pas une distinction nette des valeurs dans ce genre de graphe. La Figure 14 montre les diagrammes en radar

pour les six locutrices. Nous remarquons que les valeurs de HNR se ressemblent de manière plus marquées que le reste des variables sauf pour une locutrice qui présente une valeur de la moyenne légèrement moins élevée (diagramme à droite de la première ligne).

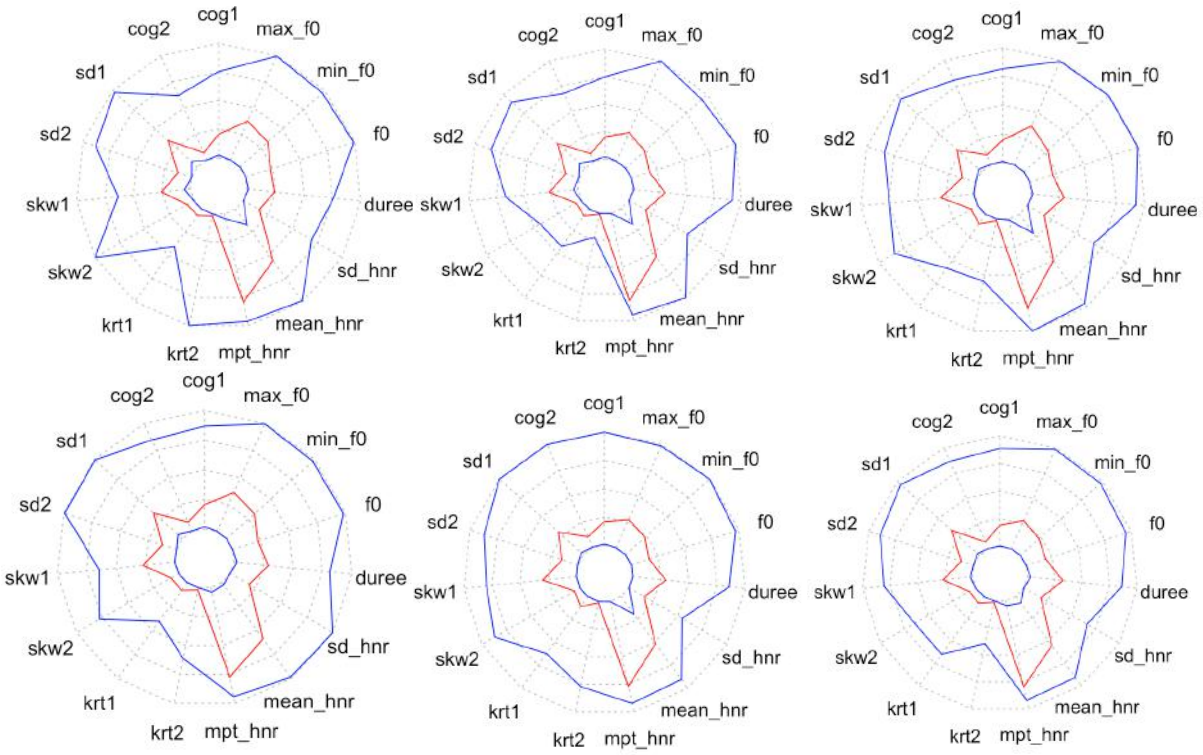

Fig. 14 diagrammes en radar des indices acoustiques pour les six locutrices du corpus PATATRA

Ceci avait été remarqué aussi au niveau interphonémique de notre analyse à travers la Figure 5 qui montrait une variation de cette mesure chez deux locuteurs du corpus sur la voyelle nasale /ã/. La tendance se confirme pour l'instant chez une locutrice mais nous confirme aussi le peu d'influence de cet indice ayant comme but primaire l'analyse de dysfonctionnements des plis vocaux sur des locuteurs sains. Cependant les graphes prennent des formes assez différentes lorsque nous regardons les autres variables, en particulier les indices représentant le kurtosis et skewness en puissance 2 et les différentes mesures de la fréquence fondamentale. Le fait que ces deux ensembles d'indices avaient des projections similaires dans l'espace réduit de l'Analyse en Composantes Principales peut nous porter à affirmer qu'il s'agit d'indices peu influencés par les altérations des représentations des données et par conséquent très robustes dans le maintien de l'information sur le locuteur. Nous allons analyser en dernier le même jeu de représentation

pour les quatre locuteurs hommes du corpus PATATRA, nous verrons si les mêmes indices jouent un rôle discriminant pour les deux sexes.

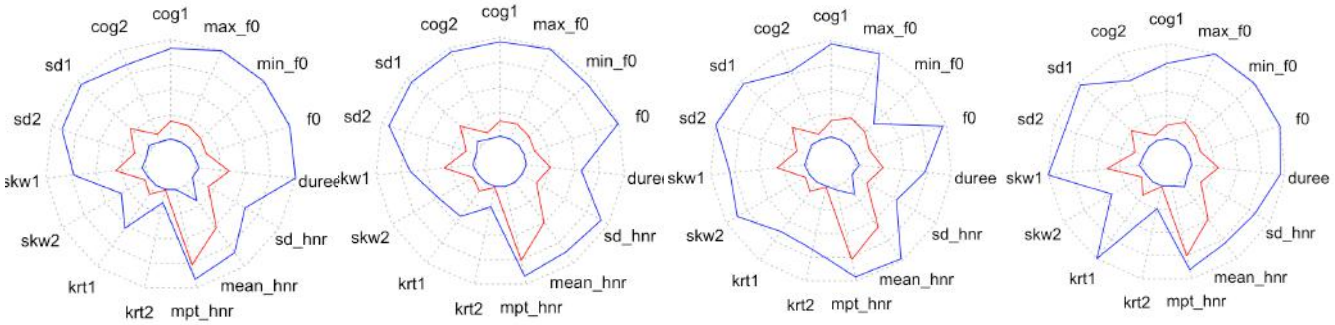

Fig. 15 diagrammes en radar des indices acoustiques pour les quatre locuteurs du corpus PATATRA

Dans ce deuxième sous-ensemble, nous retrouvons le deuxième locuteur ayant une valeur moyenne de HNR moins élevée que les autres (Figure 13b, diagramme 3) et des indices spectraux très constants. La durée constitue chez les hommes un indice plus variable que chez les locutrices où nous pouvions observer une valeur moins importante chez une seule des locutrices. Dans les diagrammes de nos quatre locuteurs l'indice de durée des segments constitue une variable discriminante avec des valeurs différentes dans chaque représentation : nous remarquons aussi que le locuteur ayant la valeur de durée la plus élevée correspond bien au locuteur 3 projeté dans l'espace de l'ACP en Figure 12 et qui se distinguait déjà pour cette valeur. Les indices de fréquence fondamentale chez les hommes semblent varier en mesure inférieure, la valeur maximale étant la mesure la plus variable les deux autres valeurs, l'estimation moyenne et la valeur minimale, restent assez constantes dans les quatre diagrammes.

Nous avons pu observer la variabilité de certains indices acoustiques porteurs d'informations sur le locuteur : nous remarquons que indépendamment du sexe la valeur maximale de la fréquence fondamentale est un indicateur fort pour la discrimination inter locuteur ; certains indices spectraux, skewness et kurtosis, se trouvent être plus discriminants pour les locutrices alors que la durée est un facteur important chez les hommes ; la moyenne de la valeur de HNR chez des locuteurs sains peut apporter une information complémentaire minimale que nous ne pouvons pas sous-estimer et qu'il faudrait peut-être investiguer de manière plus approfondie. Ces considérations comprennent les

données issus d'enregistrements en milieu contrôlé où certains indices pourraient apparaître plus ou moins évidents que d'autre dû à la situation moins naturelle dans laquelle les locuteurs ont été contraints d'enregistrer mais qui nous fournissent des extraits plus propres et précis. Dans le prochaine paragraphe nous passerons en revue les représentations des valeurs obtenues à partir de données non contrôlées et nous pourrons déduire les avantages et les désavantages des deux contextes.

6.2 FABIOLE

La base FABIOLE, le deuxième corpus de notre étude, se compose d'enregistrements de 30 locuteurs hommes avec extraits d'émissions télévisées et radiophoniques. Comme pour le corpus précédent, nous avons procédé initialement à une Analyse en Composantes Principales et nous allons discuter ici les résultats obtenus, avec cette représentation, de l'espace réduit aux deux premières composantes. La représentation par Composantes Principales des données brutes sera discutée en premier, dans ce cas nous observons une légère augmentation du pourcentage de la variabilité représentée par cet espace par rapport aux données de la Figure 11 avec un taux global de 55.14%. Les deux premières Composantes sont celles qui représentent la majorité de la variabilité, les restantes ne dépassant pas le 10% montrent des espace où le nuage des individus est très compacte et ne présente pas de valeurs qui s'éloignent du cœur.

Les données brutes présentées dans la Figure 16 se rapprochent des considérations faite pour le corpus PATATRA : nous observons un ensemble de valeurs peu distinguées qui se regroupent à l'origine de l'espace et une population inférieure qui s'écarte du centre et qui aident à expliquer la variabilité représentée par ces Composantes. Les variables projetés restent aussi cohérentes avec les autres projections analysées. Nous observons une proportionnalité inversée entre les deux moments spectraux pour les couples centre de gravité et écart type s'opposant à skewness et kurtosis. La durée a un impact mineur, son vecteur est très réduit : ceci nous amène à l'hypothèse que cet indice ne représente pas un facteur déterminant lors de la caractérisation des locuteurs pour le corpus non contrôlé. Les indices qui se trouvent être plus déterminant selon cette représentation correspondent au coefficient spectraux d'asymétrie et d'aplatissement, skewness et kurtosis. Cela revient à ce

qui avait été dit dans le premier chapitre de cette analyse sur l'influence de la forme du spectre pour la recherche d'indice de variabilité inter locuteur : pour identifier des indices plus fiables les simples mesures des pics spectraux semblent pouvoir perdre des informations déterminantes pour la caractérisation du signal sonore propre aux locuteurs, inversement des computations sur la forme générale du spectre représentent de manière fidèle les indices recherchés et qu'il faut exploiter au maximum (Blumstein et Stevens, 1981).

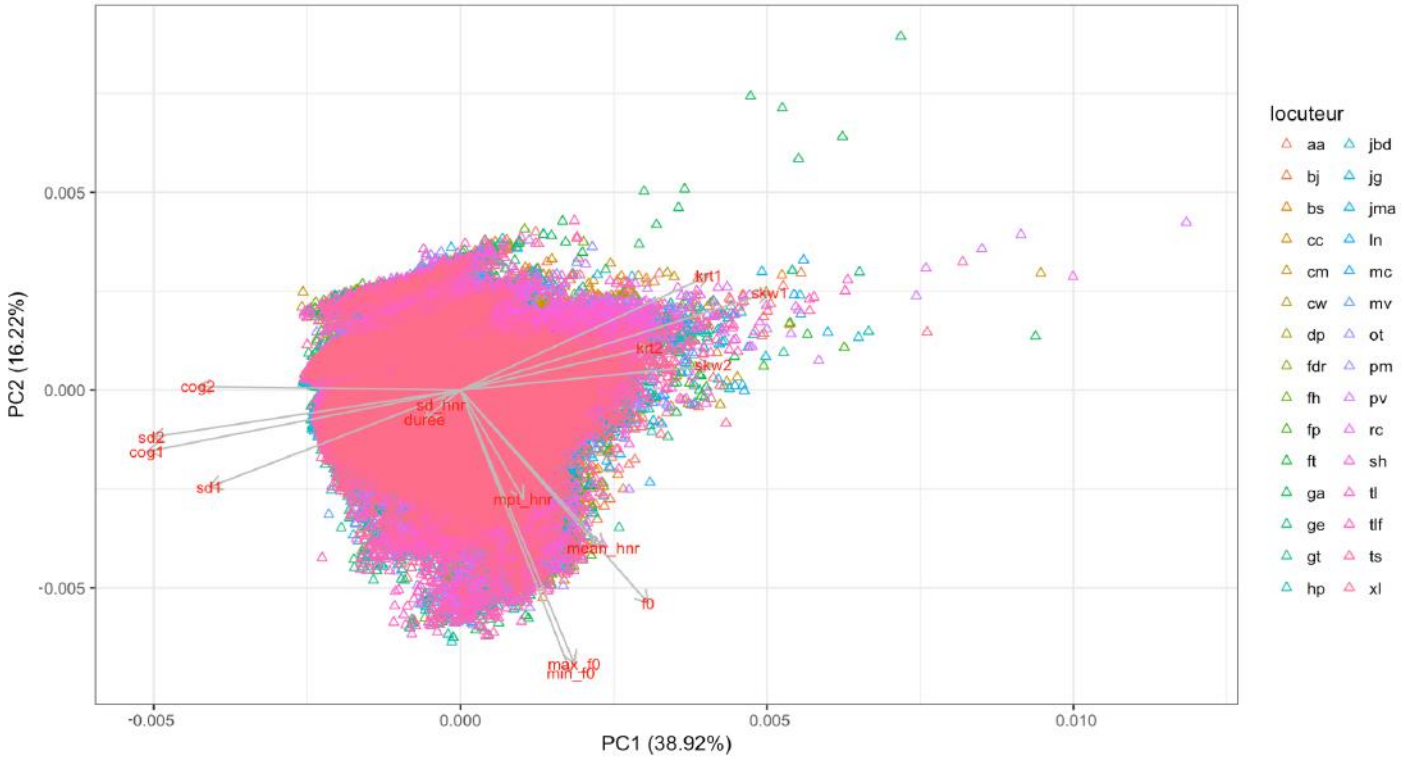

Fig. 16 représentation par Composantes Principales 1 et 2 pour du corpus FABIOLE : données brutes ; nuage des variables et des individus simultanément

De la même manière que dans la Figure 12 les locuteurs seront représentés dans l'ACP du corpus normalisé en Figure 17 par un chiffre. Cet espace réduit explique 70% de la variabilité de la population pour le corpus de 10 locuteurs en milieu contrôlé. Dans le nouveau contexte que nous analysons, ce taux s'élève à 75.62% : résultant de la combinaison de 48.89% grâce à la Composante Principale 1 et de 26.73% dans la PC2. La valeur relative de la première baisse légèrement dans ce cas, nous le rappelons c'était grâce à cet axe que nous avions pu séparer les deux sexes des locuteurs. Cette distinction étant inexistante dans le corpus FABIOLE nous nous limitons à observer que la distribution des

individus se fait de manière moins homogène que prévu : autour du point 0 nous avons une population plus chevauchante qu'aux extrémités, à partir des point 0.2 et -0.1 où les locuteurs s'écartent les uns des autres. Sur l'axe des ordonnées le même phénomène de produit sans pourtant avoir des écarts plus définis entre les individus. La distribution est certainement moins claire que pour le corpus d'une taille moins importante parce que une population majeure porte un nuage d'individus plus conséquent mais la dispersion des locuteurs dans le corpus rend cette représentation lisible et permet de distinguer de manière nette la majorité des locuteurs. À la fois dans l'ensemble brut que dans celui normalisé nous observons que la majorité de la contribution est donnée par les mesures de f_0, avec l'ajout de l'indice du HNR moyen pour les données normalisées. La plupart des individus ne présentent pas de valeurs plus élevées pour des indices spécifiques, le nuage de points se positionne globalement à l'intérieur du cercle des variables sauf pour : le locuteur 4, cc, pour l'indice de kurtosis en puissance 1 ; le locuteur 10, fp, pour la valeur du HNR ; les locuteurs 17, jg et 28 tlf, pour l'écart type de cette même mesure ; le locuteur 18, jma pour les indices de fréquence fondamentale ; les locuteurs 29, ts et 21, ot montrent enfin une valeur de l'écart type spectral plus élevée respectivement pour l'indice en puissance 1 et 2. Seuls ces sept locuteurs se placent clairement au-delà des vecteurs de variable, nous nous attendons à ce que ces locuteurs se distinguent par ces mêmes indices à travers l'arbre de décision en Figure 18.

Une dernière remarque concerne la projection des variables dans l'espace, nous observons une différence évidente par rapport à celles du corpus PATATRA. Tous les indices dans cette ACP se distribuent sur un semi-diamètre alors que dans le premier nous retrouvions aussi des indices se projetant à l'extérieur de cette figure géométrique, notamment kurtosis et skewness en puissance 1, qui s'opposaient aux indices de HNR et aux trois autres moments spectraux. Cette deuxième opposition existe aussi pour les individus du corpus FABIOLE, nous en déduisons que des locuteurs présentant des indices spectraux de skewness et kurtosis plus élevés auront des valeurs moins importantes pour les autres moment spectraux. La fréquence fondamentale et la durée forment ici un angle aigu avec cette dernière qui est projetée dans la même direction que les indices de HNR alors que dans le contexte contrôlé ils se trouvaient être presque perpendiculaires. Comme pour l'ACP précédent nous nous limitons à observer les résultats des deux premières

composantes puisque les successives se limitent à décrire chacun des pourcentages très bas de la population complémentaires aux 75% des deux que nous venons d'exposer.

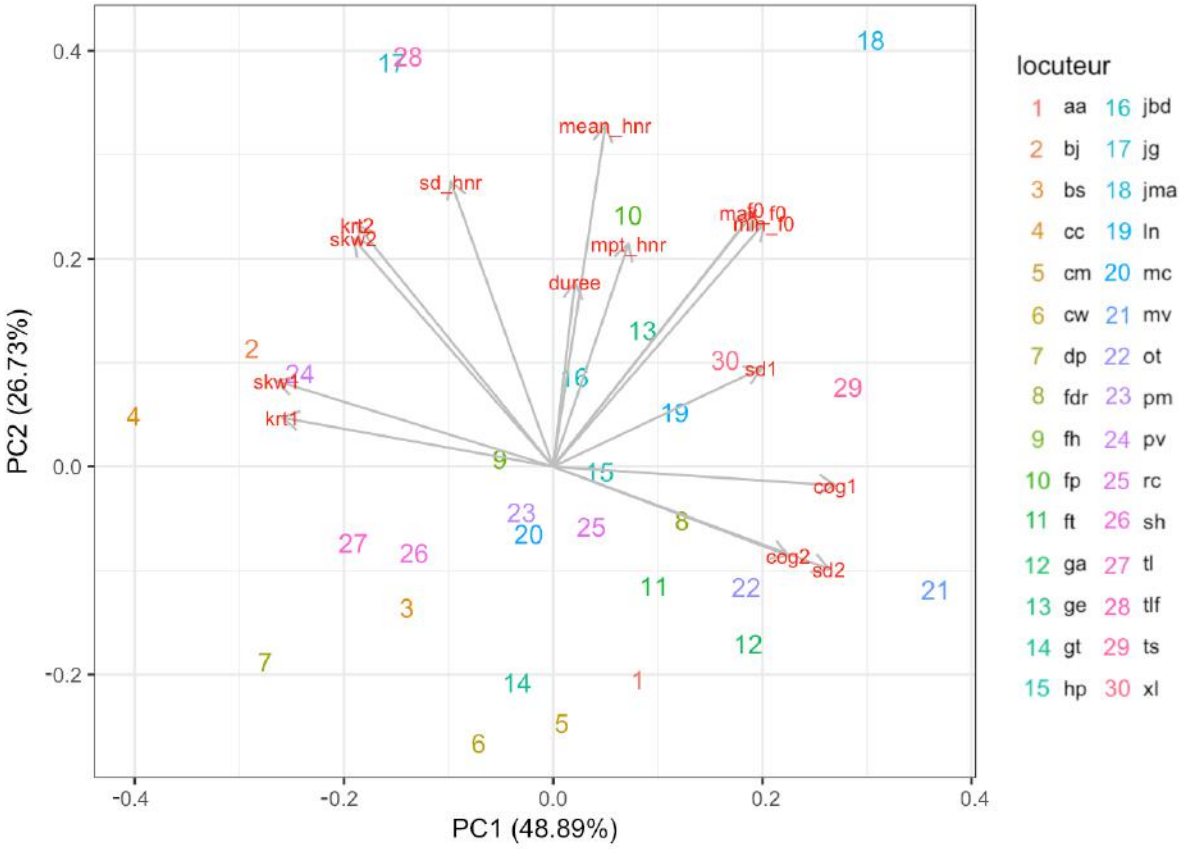

Fig. 17 représentation par Composantes Principales 1 et 2 pour le corpus FABIOLE : données normalisées ; nuage des variables et des individus simultanément

Comme pour le corpus précédent, l'arbre de classification construit pour ce corpus, Figure 18, nous avons varié l'ensemble des variables utilisées lors de la prise de décision et nous avons retenu le graphe capable de décrire la variabilité inter locuteur de la façon la plus représentative. Cependant, différemment de celui en Figure 12, l'arbre résultant du corpus FABIOLE ne décrit pas entièrement le corpus. Nous avions anticipé que ce genre de représentation est plus facilement interprétable mais présente l'inconvénient majeur d'être très influencée par la fluctuation des échantillons : notre tableau de résumé du corpus présente une taille conséquente (autour du million et demi d'observations) ainsi que des valeurs très homogènes, nous venons de voir à travers l'ACP que seuls sept locuteurs sur trente présentent des valeurs plus élevées pour des indices spécifiques. Ces éléments ont rendu la formation de notre arbre une tâche peu efficace par rapport aux attentes.

Uniquement six locuteurs sur trente arrivent à être distinguer de manière claire selon cette représentation, nous verrons si les variables permettant cette classification réduite correspondent aux indices observés dans l'espace réduit de l'ACP. La classification effectuée grâce à cet arbre distingue des locuteurs dont nous avions peu d'informations à travers l'Analyse en Composantes Principales. Il s'agit de locuteurs à la fois placés aux extrémités du nuage d'individus, le cas du locuteur 4, cc, et de locuteurs qui se trouvaient dans le cœur du nuage, notamment 8, fdr et 9, fh.

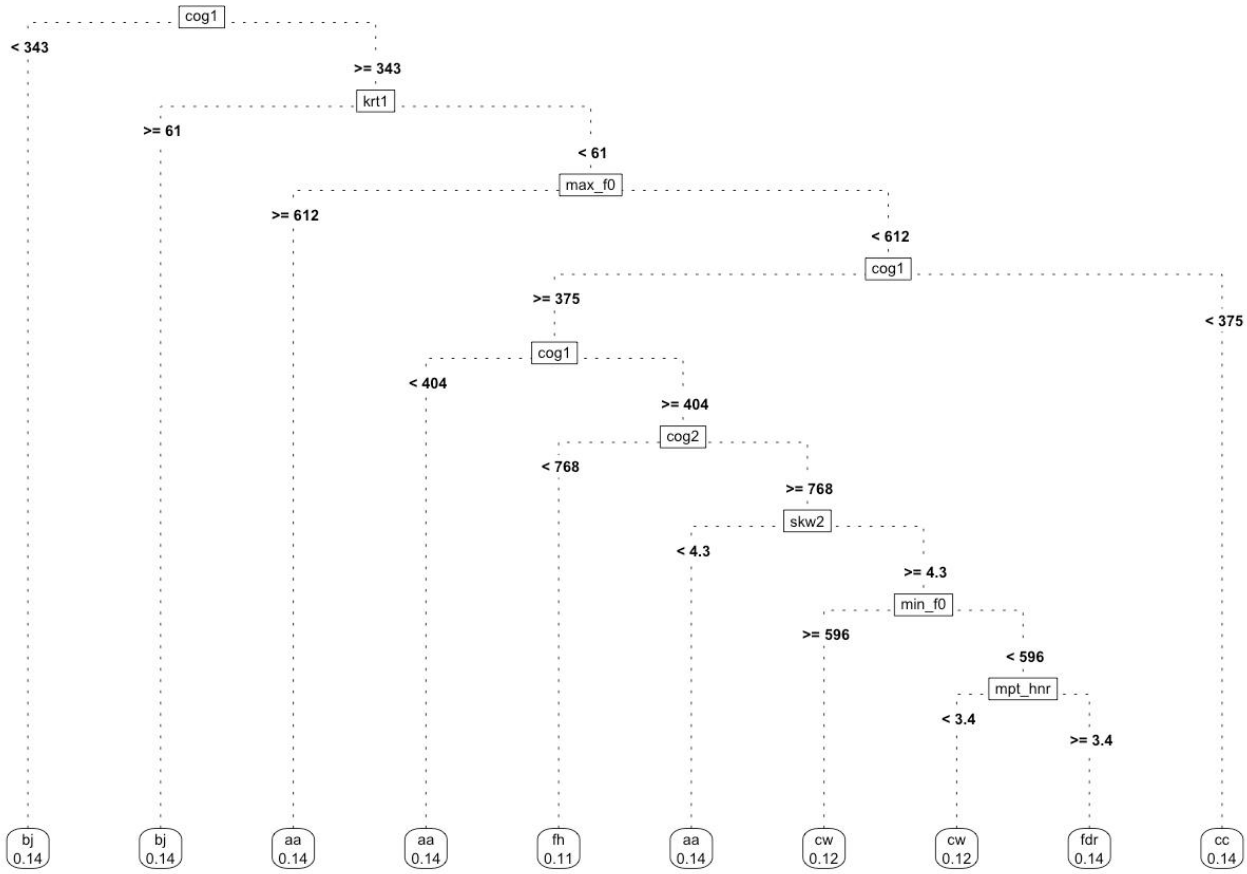

Fig. 18 arbre de classification pour le corpus FABIOLE, données quantitatives

Un élément en particulier captive notre attention si nous comparons cet arbre et celui en Figure 13 pour le corpus PATATRA : l'indice du centre de gravité spectral ainsi que les autres moments spectraux se trouvent être très discriminants dans ce cas alors que la fréquence fondamentale joue un rôle moins important apparaissant uniquement sur deux

branches. Cette majeure efficacité dans la discrimination de différents locuteurs de la part des moments spectraux pour des extraits de parole dans un contexte non contrôlé est un des points centraux de notre étude : nous l'avons observé dans l'ouvrage de (Kahn, 2011) auquel nous nous sommes confrontés et inspirés pour mener cette analyse ; nous l'avons aussi observé au niveau phonémique lors du chapitre précédent. Les deux mesures des moments spectraux, en particulier le premier moment, se trouvent être les variables porteuses de la plupart de l'information sur le locuteur alors que les indices de fréquence fondamentale et du rapport Harmoniques sur bruit sont complémentaires pour affiner la classification lorsque nous observons des données issues d'un milieu non contrôlé. Pour le contexte contrôlé pour lequel nous avons pu constater une tendance inverse avec la f_0 qui joue un rôle prépondérant alors que les moments spectraux et le HNR représentent un complément à l'information décrite par les mesures de ce premier indice.

Nous allons maintenant nous concentrer sur les diagrammes en radar des 30 locuteurs du corpus FABIOLE, nous discuterons de l'efficacité de cette représentation graphique et constaterons si d'autres éléments ressortent de cette analyse. Nous aborderons également des suggestions pour des études ultérieures. L'ensemble des trente diagrammes est en Figure 18 nous avons gardé le même style de représentation que pour les diagrammes du corpus précédent : les différents indices acoustiques aux extrémités de la toile d'araignée ; chaque locuteur ayant son propre diagramme et dessinant sa propre forme de manière plus ou moins évidente.

Les considérations faites jusqu'ici pour le corpus FABIOLE se confirment dans ces diagrammes. Globalement nous observons des motifs par locuteur assez similaires, pour les indices de HNR et de durée. Ces valeurs proches pour tous les locuteurs peuvent être dû : à l'absence de dysfonctionnements des plis vocaux chez les locuteurs pour le HNR, en s'agissant de présentateurs d'émissions ainsi que d'invités lié au monde de la politique et de la communication nous pouvons supposer qu'il s'agit en grande partie de gens qui sont porté à avoir une bonne qualité de voix dans leur occupation ; pour la durée une considération similaire peut être faite en rajoutant que les émissions ayant des temps bien précis à respecter la vitesse d'élocution serait plus importante et la durée inversement proportionnelle à celle-ci, le fait que cette valeur apparaît constante chez les 30 locuteurs confirme cette hypothèse d'homogénéité. Pour les moments spectraux nous observons des variations

moins importantes que nous ne l'attendions : le centre de gravité et le coefficient d'asymétrie du spectre en puissance 1 ainsi que l'écart type en puissance 2 sont les variables les plus touchées par la variation inter locuteur et présentent une corrélation entre elles. Lorsque nous observons une montée des valeurs du centre de gravité dans le diagramme, inversement nous observons une descente du coefficient d'asymétrie, cette affirmation est valable aussi pour l'écart type. Ceci confirme la tendance que nous avions observé lors de l'ACP où l'écart type et le centre de gravité ayant une projection similaire s'opposaient à l'asymétrie. Les mesures de fréquence fondamentale montrent également des variation dans ces diagrammes qui se traduisent en une discrimination importante des locuteurs entre eux à travers cet indice.

Dans l'ensemble des observations que nous avons pu faire jusque-là nous remarquons que dans un contexte où l'élocution est moins contrôlée, mais toutefois contrainte par des facteurs autres que ceux de l'environnement d'enregistrement, les éléments de variabilité inter locuteur se rapprochent de ceux étudiés en milieu contrôlé. La fréquence fondamentale joue toujours un rôle important avec la valeur maximale qui parmi les trois mesures étudiées (estimation moyenne, valeur minimale et valeur maximale), peut être considérée la variable la plus discriminante puisqu'elle agit dans la majorité des classifications. Les mesures spectrales sont plus porteuses d'informations lorsque le milieu d'enregistrement est moins contrôlé, elles se distinguent comme variables complémentaires dans l'autre contexte analysé et restituent des informations sur le caractère statique du signal sonore. Une exploration plus approfondie des indices spectraux pourrait porter à l'identification d'autres informations dans le signal. Le HNR s'est révélé être un complément sur l'information également dans un contexte de voix moins contrôlé lors de la tâche de classification des locuteurs selon les valeurs des indices acoustiques analysés lors de cette étude.

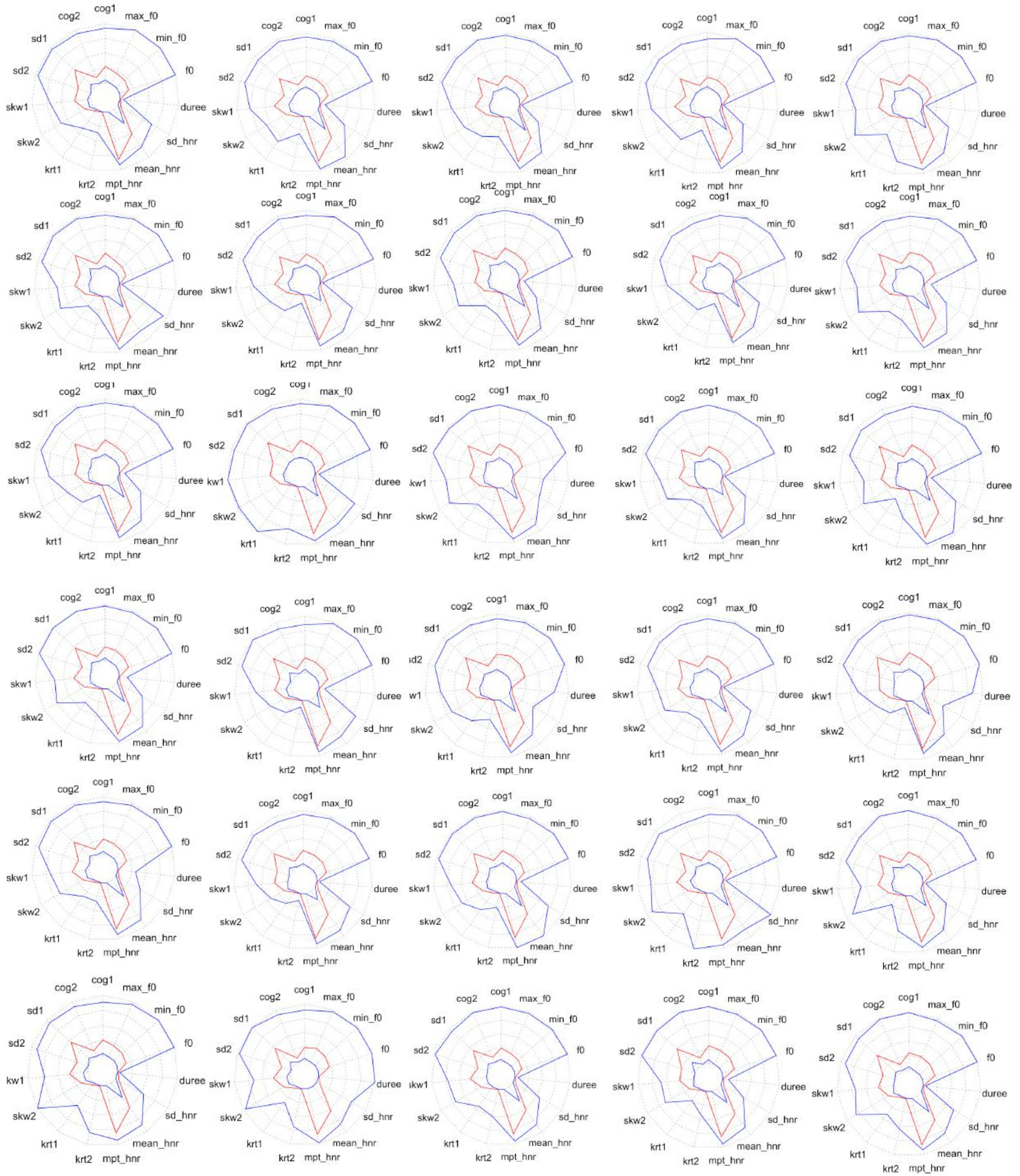

Fig. 19 diagrammes en radar des indices acoustiques pour les trente locuteurs du corpus FABIOLE

6.3 MFCC

Nous avons procédé aux représentations par ACP, arbre de classification et diagrammes en radar également pour les données des coefficients cepstraux des deux corpus : l'intérêt étant de comprendre en quelle mesure ces indices habituellement utilisés en Reconnaissance Automatique du Locuteur se comportent lorsque nous analysons les données de manière plus approfondie. Après l'analyses des résultats de ces trois méthodes, nous avons décidé de présenter uniquement les représentations dans l'espaces des Composantes Principales car les deux autres résultats obtenus ne se révèlent pas être suffisamment discriminants au niveau inter locuteur.

L'arbre de classification construit à travers les coefficients cepstraux n'arrive à dissocier que quatre locuteurs sur dix pour les données du corpus PATATRA et uniquement six locuteurs sur trente pour le corpus FABIOLE. La combinaison de valeurs des 13 coefficients cepstraux que nous avons calculée semble ne pas être déterminant pour une tâche de classification qui considère ces valeurs de manière séparée. Pour les diagrammes en radar nous assistons a des motifs presque identiques : nous considérons l'hypothèse que les valeurs représentées par ce diagrammes ne montrent qu'une caractéristique statique du locuteur alors que les coefficients cepstraux effectuent une description dynamique du signal sonore.

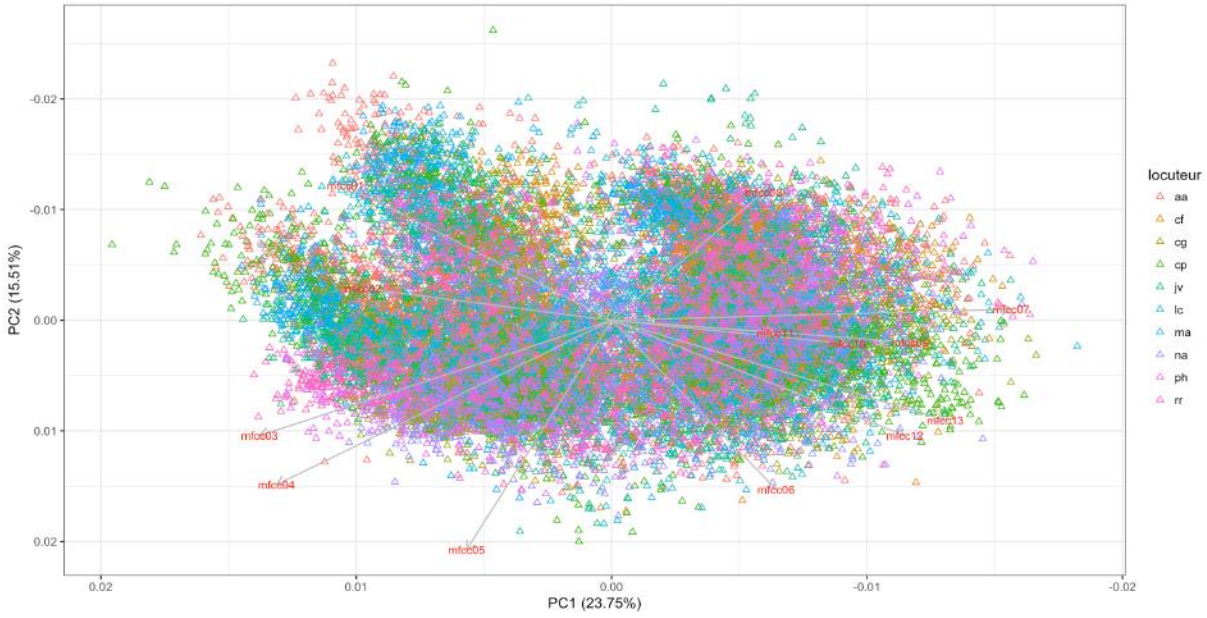

Fig. 20 représentation par Composantes Principales 1 et 2 des MFCC du corpus PATATRA : données brutes ; nuage des variables et des individus simultanément

Ces deux représentations ayant montré des points faibles, nous nous concentrons sur l'ACP qui, quant à elle, devrait donner une interprétation plus discriminante des données.

Dans les Figure 20 et 21 nous observons que les individus pour les données issues du corpus PATATRA et FABIOLE respectivement, montrent encore une fois une distribution très compacte avec certaines valeurs qui s'éloignent du corps du nuage. Nous ne pouvons pas distinguer nettement les valeurs d'un locuteur par rapport à un autre mais grâce aux projections des variables, les 13 coefficients cepstraux dans ce cas, nous avons des indications sur leur impact dans la caractérisation des locuteurs. Pour le corpus en milieu contrôlé cette transposition dans l'espace réduit arrive à expliquer une variabilité combinée de 39.26%, ce qui est nettement en dessous des valeurs montrées par les indices acoustiques. Les deux premières Composantes pour le corpus FABIOLE montrent un pourcentage encore moins important de variabilité avec une valeur combinée de 29.47%.

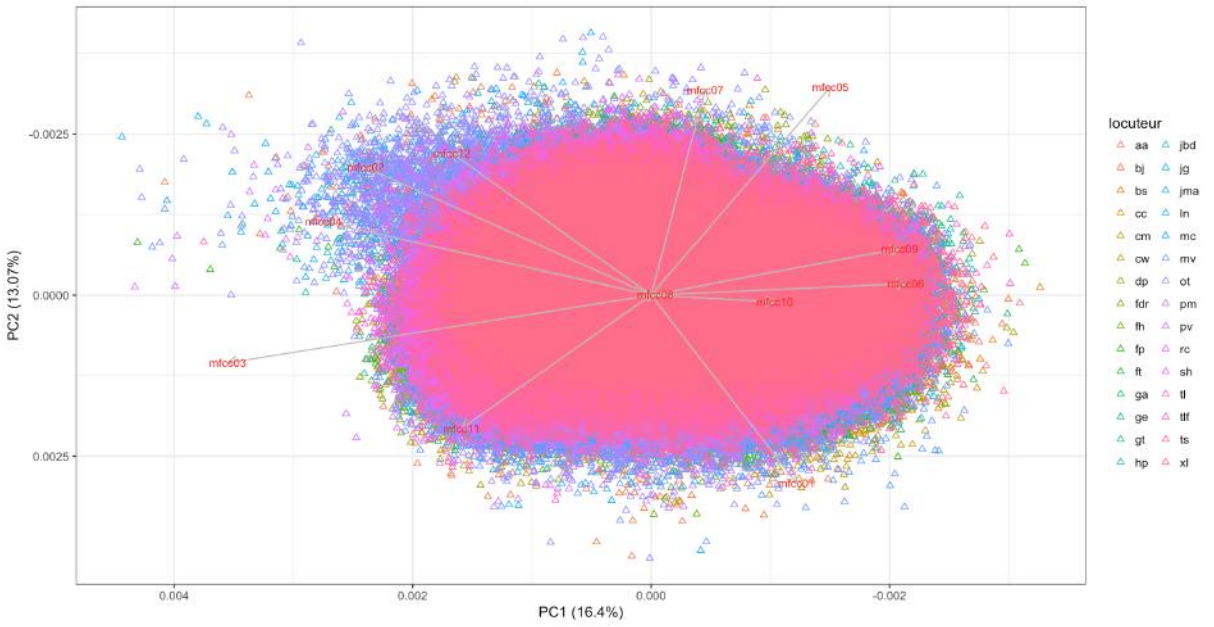

Fig. 21 représentation par Composantes Principales 1 et 2 des MFCC du corpus FABIOLE : données brutes ; nuage des variables et des individus simultanément

Ce que nous observons dans les deux ACP rejoint ce que nous avons pu observer dans le chapitre précédent avec certains coefficients qui se montrent comme plus porteurs d'informations sur le locuteur. Ces coefficients n'étaient pas les mêmes pour les deux corpus au niveau phonémique et ceci se répète lors de cette nouvelle analyse. Nous observons que

pour le corpus PATATRA les coefficients avec le pourcentage de contribution plus élevé sont les numéros 5, 4, 3 et 6 avec une contribution cumulée de 57%. Les coefficients 5 et 3 se trouvent être contributeur aussi pour le corpus FABIOLE, à ces deux s'ajoutent les 2 et 7. Un élément remarquable est l'absence totale de contributions dans cet ensemble pour les coefficients 1 et 13. La différente influence qui ont les coefficients cepstraux sur les deux corpus est un élément qui est revenu dans notre analyse mais sur lequel nous n'avons pas d'explication précise. Les tests que nous avons effectué nous ont donné des indications de départ et il faudra les approfondir dans l'avenir pour comprendre pleinement la contribution que ces indices ont dans la caractérisation du locuteur. Les MFCC constituent un ensemble de paramètres, comme les mesures acoustiques que nous avons discuté, qui sont porteurs à la fois d'informations majeures sur les locuteurs ainsi que d'informations complémentaires pour la discrimination. L'interprétation de ces paramètres reste, cependant, pas simple puisqu'ils ne relèvent ni de caractéristiques physiologiques, l'exemple du HNR, ni d'informations spectrales comme dans le cas des moments spectraux. La contribution des coefficients cepstraux dans le domaine de la Reconnaissance Automatique du Locuteur est toutefois indéniable mais nous avons pu observer comment ils perdent en robustesse lors de l'analyse critique des valeurs caractérisant les locuteurs. La recherche de techniques capables d'expliquer de manière plus exhaustive en quelle mesure ces facteurs agissent pour la caractérisation des locuteur, nous avons montré les avantages et désavantages de trois différentes techniques de représentation, est un point majeur de réflexion pour la suite des travaux d'analyse et de comparaison entre des indices acoustiques classiques et les paramètres utilisés en RAL.

7 Conclusion et perspectives

L'étude que nous avons présentée ici n'est qu'une première exploration de nouveaux indices acoustiques qui pourraient porter davantage d'informations sur le locuteur lors des tâches de reconnaissance. La limite principale de notre étude, mais aussi le point d'amélioration le plus important est représentée par le corpus en milieu contrôlé. Les indices acoustiques montrent une majeure efficacité avec des enregistrements de ce genre mais notre corpus étant limité à 10 locuteurs (6 femmes et 4 hommes) numériquement il ne peut pas représenter un échantillon assez représentatif de la population. Cependant, l'utilisation d'une plus large base de voix contrôlées pour une phase d'apprentissage ou pour continuer à explorer l'analyse d'indices pourrait représenter une particularité sur la même échelle que les croisements entre langues d'apprentissage et de test pour les systèmes de reconnaissance de la langue que nous retrouvons dernièrement dans la littérature de ce domaine (Shon et al., 2017 ; Dunbar et al., 2017). Ces mêmes auteurs insistent sur l'utilisation d'indices phonétiques ou phonémiques propres à la langue parlée par le locuteur pour améliorer les systèmes de reconnaissance les rendant des systèmes hybrides langue-locuteur : ceci se rapprocherait de la façon dont le cerveau humain pratique la reconnaissance, de manière conjointe. Cette idée de fond de relier les deux domaines pourrait être poussée.

Une partie de la littérature sur la reconnaissance du locuteur se concentre dernièrement également sur l'utilisation de segments ne correspondant pas directement à des sons linguistiques (Dumpala et al., 2017). Par exemple, nous retrouvons l'analyse des pauses ou du rire qui reviennent dans (Zhang et al., 2017b ; Zhao et al., 2017), une analyse plus approfondie de ces éléments particuliers à l'intérieur du discours pourrait se révéler porteuse de nombreuses informations sur les locuteurs.

Parmi les mesures que nous avons pu montrer, la fréquence fondamentale et les mesures de rythme peuvent être considérées comme étant une voie à explorer de plus près pour de futures analyses, en prenant des mesures de rythme autres que les PVI ou des calculs plus précis sur la dynamique de la f_0. Les moments spectraux ont aussi un rôle important dans notre analyse, ils ont montré des améliorations intéressantes grâce aux deux différentes méthodes de calcul, d'autres valeurs de puissance pourraient être évaluées ou encore

différentes représentations du spectre. Les coefficients cepstraux, au centre des systèmes de RAL et analysés dans notre cas de manière statique, ont montré des limites par rapport à d'autres indices acoustiques, notamment pour les fricatives et les occlusives. Sur ces deux catégories de segments des mesures de coarticulation pourraient faire l'objet d'analyses ultérieures, à travers l'utilisation de bigrammes.

Annexes

Annexe 1 : déclaration sur l'honneur

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

Date : 2/7/2018

Signature de l'étudiant

Annexe 2 : valeurs de la taille d'effet pour les deux corpus

Valeurs de η2 - corpus Fabiole

MS1.1	**p**	**t**	**k**	**b**	**d**	**m**
	0.06152117	0.02713769	0.01586122	0.2090545	0.1453507	0.3162885
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.1429962	0.08801194	0.08935277	0.1525715	0.1364876	0.0922074
	ɑ̃	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.3384345	0.3701232	0.2538888	0.3071915	0.1487939	0.09259048
MS1.2	**p**	**t**	**k**	**b**	**d**	**m**
	0.02109349	0.01661559	0.01529154	0.1775272	0.1542157	0.1938105
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.1485437	0.06710779	0.08413137	0.1042678	0.1165762	0.07317953
	ɑ̃	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.2791708	0.3260365	0.1684466	0.2785808	0.1077902	0.0278388
MS2.1	**p**	**t**	**k**	**b**	**d**	**m**
	0.08163774	0.05352834	0.06518343	0.1799511	0.1346561	0.2621354
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.09252716	0.08847849	0.1322172	0.2149997	0.1309546	0.1028668
	ɑ̃	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.2384865	0.2518671	0.2135583	0.164796	0.1420826	0.133901
MS2.2	**p**	**t**	**k**	**b**	**d**	**m**
	0.04598523	0.02044781	0.01330589	0.1944795	0.1421147	0.2742008
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.1475265	0.06355483	0.08934495	0.1614602	0.1111773	0.09616111
	ɑ̃	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.3067008	0.3386832	0.1920469	0.2572496	0.1290126	0.05310974
MS3.1	**p**	**t**	**k**	**b**	**d**	**m**
	0.09084223	0.02894094	0.02090505	0.1859875	0.1331756	0.3174093
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.1268292	0.1092147	0.07126988	0.1779676	0.1448033	0.07095259
	ɑ̃	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**

	0.3280212	0.3248788	0.2750253	0.2445254	0.1555115	0.1178939
MS3.2	**p**	**t**	**k**	**b**	**d**	**m**
	0.06632137	0.0419001	0.04027957	0.08129637	0.1176945	0.200759
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.1534808	0.06723418	0.06473582	0.06641354	0.1255696	0.09528364
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.1440177	0.2322053	0.1830671	0.1807141	0.178343	0.07650213
MS4.1	**p**	**t**	**k**	**b**	**d**	**m**
	0.09394323	0.02302996	0.02783142	0.1593735	0.109702	0.287269
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.09600292	0.1107107	0.05643321	0.152967	0.1494502	0.03603431
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.2867203	0.2938324	0.2506756	0.2043615	0.1156772	0.1244363
MS4.2	**p**	**t**	**k**	**b**	**d**	**m**
	0.0435039	0.01718805	0.02150268	0.04970327	0.04499031	0.1707354
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.1016382	0.05187259	0.03397896	0.0169609	0.1012672	0.03421792
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.1839114	0.151218	0.2050076	0.2018332	0.09934171	0.05174247
HNR mid	**p**	**t**	**k**	**b**	**d**	**m**
	0.02792956	0.03864596	0.04473671	0.01095424	0.00899607	0.01359954
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.00926422	0.0167387	0.01042917	0.06511483	0.02899886	0.02995304
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.02145741	0.04284771	0.0810535	0.01486593	0.02563193	0.02076872
HNR mean	**p**	**t**	**k**	**b**	**d**	**m**
	0.07233613	0.04497185	0.03694193	0.07991481	0.09144513	0.07792691
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.06862253	0.04683365	0.08656127	0.1170072	0.09309051	0.09953218
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.07923702	0.107509	0.1110508	0.07980562	0.04973569	0.06558091

HNR sd	p	t	k	b	d	m
	0.02251226	0.01972329	0.03015524	0.02798904	0.01640956	0.02692555
	l	ʁ	f	ʃ	v	z
	0.01268714	0.00844191	0.07296842	0.08038189	0.01756845	0.04657123
	ã	ɛ̃	ɔ̃	a	i	u
	0.01947363	0.01664391	0.01579306	0.01650777	0.04063109	0.01482707
f_0	p	t	k	b	d	m
	0.07720556	0.07941087	0.06646548	0.2265897	0.2263648	0.2592355
	l	ʁ	f	ʃ	v	z
	0.2411319	0.09793009	0.2502896	0.2429751	0.2310285	0.1831466
	ã	ɛ̃	ɔ̃	a	i	u
	0.2042863	0.2303143	0.2081275	0.1167824	0.21233	0.2006134
f_0 min	p	t	k	b	d	m
	0.06889002	0.07477304	0.05503961	0.2137024	0.2276479	0.2690939
	l	ʁ	f	ʃ	v	z
	0.2296823	0.08437856	0.09243096	0.08146629	0.2353758	0.1811693
	ã	ɛ̃	ɔ̃	a	i	u
	0.2161363	0.2372619	0.2245566	0.1168826	0.2197842	0.2048198
f_0 max	p	t	k	b	d	m
	0.08624512	0.09055109	0.04730201	0.1630848	0.1945199	0.2242211
	l	ʁ	f	ʃ	v	z
	0.1719309	0.07537082	0.1225032	0.07255166	0.2056371	0.166043
	ã	ɛ̃	ɔ̃	a	i	u
	0.1548244	0.1756685	0.1598937	0.07681027	0.1650312	0.159544
Durée	p	t	k	b	d	m
	0.03985692	0.02746895	0.03150997	0.03330088	0.02097519	0.04063392
	l	ʁ	f	ʃ	v	z
	0.02037888	0.02992919	0.04149333	0.0639051	0.03689382	0.02508599
	ã	ɛ̃	ɔ̃	a	i	u
	0.04675062	0.0390024	0.03642305	0.0216016	0.02025634	0.01494997
MFCC1	p	t	k	b	d	m

	0.06757912	0.0398965	0.0398016	0.198857	0.1259684	0.2252711
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.06093266	0.06221734	0.06607145	0.09114997	0.09589584	0.05321375
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.3459085	0.359632	0.2820994	0.2929694	0.1085949	0.1293459
MFCC2	**p**	**t**	**k**	**b**	**d**	**m**
	0.2088179	0.1574548	0.1231494	0.2209232	0.2123661	0.2842963
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.155361	0.08552797	0.1356936	0.1986825	0.1466449	0.08194653
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.3202758	0.296345	0.3354261	0.2382225	0.1677763	0.1326137
MFCC3	**p**	**t**	**k**	**b**	**d**	**m**
	0.1545679	0.1460063	0.1166896	0.1057754	0.1247904	0.2008726
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.09163981	0.1224586	0.1492342	0.2534923	0.1091674	0.1169115
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.2944678	0.2393398	0.2755197	0.2067596	0.2041732	0.1636
MFCC4	**p**	**t**	**k**	**b**	**d**	**m**
	0.1364642	0.1323718	0.1084641	0.1128746	0.1496319	0.2438308
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.09543247	0.07900409	0.1123271	0.1775119	0.07700775	0.1017647
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.2805329	0.309768	0.2709498	0.2467832	0.2026953	0.09437785
MFCC5	**p**	**t**	**k**	**b**	**d**	**m**
	0.09846959	0.1026007	0.09186232	0.1200664	0.1036457	0.3313536
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.1397857	0.09175107	0.1195276	0.1733914	0.1272835	0.1241923
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.2435939	0.3391188	0.2283558	0.2505818	0.2493289	0.1575249
MFCC6	**p**	**t**	**k**	**b**	**d**	**m**
	0.09191682	0.08631977	0.07095223	0.07748407	0.08838155	0.2364618

	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.06011335	0.08730048	0.0975802	0.1996944	0.07594358	0.1149054
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.3285306	0.2490556	0.3068833	0.2153628	0.1776824	0.1533559
MFCC7	**p**	**t**	**k**	**b**	**d**	**m**
	0.08606806	0.07700743	0.06771929	0.1195558	0.09796194	0.3373266
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.09515595	0.08281807	0.09095507	0.1135607	0.1192978	0.09023261
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.4198313	0.349642	0.3842565	0.2820438	0.1941214	0.1234035
MFCC8	**p**	**t**	**k**	**b**	**d**	**m**
	0.03580666	0.03005309	0.03242711	0.1012289	0.06735169	0.2865822
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.07358083	0.07691354	0.04667024	0.09332543	0.06532171	0.0460487
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.2651114	0.1815938	0.2940691	0.1407065	0.1901632	0.1340243
MFCC9	**p**	**t**	**k**	**b**	**d**	**m**
	0.02650484	0.03177736	0.03422262	0.05322406	0.06208842	0.1408669
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.0706216	0.07669076	0.04198585	0.09055665	0.07464522	0.06617373
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.2647272	0.3160842	0.237414	0.2304926	0.09223061	0.0942307
MFCC10	**p**	**t**	**k**	**b**	**d**	**m**
	0.03131724	0.02145193	0.03051633	0.06430269	0.05264408	0.1902358
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.08571702	0.07030893	0.02718115	0.08805248	0.04876673	0.04762878
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.2535231	0.288675	0.2858267	0.1915091	0.1560828	0.06451525
MFCC11	**p**	**t**	**k**	**b**	**d**	**m**
	0.03552377	0.02983793	0.03851112	0.07381312	0.0543565	0.2372037
	l	**ʁ**	**f**	**ʃ**	**v**	**z**

	0.08699508	0.08559119	0.04439519	0.04215469	0.0858175	0.04858574
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.2993672	0.1769783	0.2752638	0.1909362	0.1143585	0.1086977
MFCC12	**p**	**t**	**k**	**b**	**d**	**m**
	0.05081981	0.05127006	0.05125406	0.1030296	0.08413374	0.2247281
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.0907017	0.1053265	0.06440417	0.05260032	0.1137543	0.03960966
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.2112313	0.2421928	0.2254861	0.1973952	0.1655007	0.1392491
MFCC13	**p**	**t**	**k**	**b**	**d**	**m**
	0.0196819	0.02163798	0.02017162	0.02551394	0.03134208	0.1487156
	l	**ʁ**	**f**	**ʃ**	**v**	**z**
	0.04020296	0.05691503	0.03398743	0.06562279	0.05134957	0.08044575
	ã	**ɛ̃**	**ɔ̃**	**a**	**i**	**u**
	0.1547091	0.1637247	0.152115	0.1108983	0.06872436	0.05678751

Valeurs de η2 - corpus η2 Patatra, global

MS1.1	**p**	**t**	**k**	**b**	**d**	**l**
	0.133587	0.04136968	0.07386103	0.146123	0.2743029	0.4504757
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.3231431	0.434103	0.4179154	0.2882838	0.5497378	0.3990579
	m	**ã**	**a**	**i**	**u**	
	0.4661625	0.6198855	0.5506758	0.4717788	0.4040213	
MS1.2	**p**	**t**	**k**	**b**	**d**	**l**
	0.1408032	0.122698	0.06400247	0.2118239	0.07845248	0.3979892
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.3731709	0.3540468	0.4741114	0.1063063	0.3578331	0.1830519
	m	**ã**	**a**	**i**	**u**	
	0.7139736	0.5076198	0.4298568	0.148719	0.4726293	
MS2.1	**p**	**t**	**k**	**b**	**d**	**l**
	0.05862674	0.04911907	0.1182691	0.1699942	0.2500255	0.377035

	ʁ	f	ʃ	v	z	ʒ
	0.1868089	0.3343186	0.4778256	0.326071	0.2812397	0.4042628
	m	ã	a	i	u	
	0.4456828	0.6121881	0.416879	0.4464887	0.4438562	
MS2.2	p	t	k	b	d	l
	0.1564886	0.02682187	0.04973797	0.07954702	0.1489397	0.2117484
	ʁ	f	ʃ	v	z	ʒ
	0.3486294	0.3121853	0.3970476	0.1378234	0.4876759	0.2767712
	m	ã	a	i	u	
	0.2056988	0.5118825	0.4173099	0.2666428	0.04607808	
MS3.1	p	t	k	b	d	l
	0.07635769	0.05845549	0.1171363	0.1461897	0.3533931	0.4310487
	ʁ	f	ʃ	v	z	ʒ
	0.2360128	0.4215307	0.4798771	0.3056505	0.4541873	0.3801484
	m	ã	a	i	u	
	0.4312213	0.6812935	0.5172932	0.4664029	0.5213116	
MS3.2	p	t	k	b	d	l
	0.08732519	0.05202347	0.02217969	0.199208	0.2546114	0.2038232
	ʁ	f	ʃ	v	z	ʒ
	0.1784879	0.07832702	0.2049398	0.2189025	0.3379314	0.3447419
	m	ã	a	i	u	
	0.335046	0.4449871	0.254328	0.3353041	0.3031612	
MS4.1	p	t	k	b	d	l
	0.05009499	0.04949588	0.1356036	0.1447884	0.3260631	0.3986419
	ʁ	f	ʃ	v	z	ʒ
	0.1501604	0.2056923	0.5663606	0.2594868	0.1819769	0.4190297
	m	ã	a	i	u	
	0.4160311	0.6769724	0.4924694	0.4547848	0.5335044	
MS4.2	p	t	k	b	d	l
	0.06902309	0.02735695	0.02399755	0.1720946	0.1782635	0.129642
	ʁ	f	ʃ	v	z	ʒ

	0.06827977	0.01011736	0.09507488	0.1534455	0.1658338	0.1825233
	m	**ã**	**a**	**i**	**u**	
	0.2026991	0.2594934	0.1957948	0.3165935	0.1962261	
HNR mid	**p**	**t**	**k**	**b**	**d**	**l**
	0.08404637	0.08761852	0.05768452	0.05021085	0.06184763	0.09654587
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.05527464	0.03949385	0.243661	0.08639614	0.0650597	0.2502499
	m	**ã**	**a**	**i**	**u**	
	0.3983422	0.6296382	0.2266835	0.1356712	0.2493735	
HNR mean	**p**	**t**	**k**	**b**	**d**	**l**
	0.06678073	0.1168592	0.02746824	0.3323985	0.3142058	0.3309667
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.1167347	0.1289125	0.3263845	0.2815398	0.2749215	0.258525
	m	**ã**	**a**	**i**	**u**	
	0.4932286	0.709903	0.3054859	0.2547371	0.321255	
HNR sd	**p**	**t**	**k**	**b**	**d**	**l**
	0.06161836	0.07752766	0.04468719	0.1601378	0.1795504	0.06941288
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.08930604	0.1070689	0.09554706	0.1107008	0.1553718	0.1456183
	m	**ã**	**a**	**i**	**u**	
	0.3498628	0.262117	0.09814966	0.04927654	0.05772426	
f_0	**p**	**t**	**k**	**b**	**d**	**l**
	/	/	/	0.8306317	0.8000203	0.6049002
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.2937731	0.4505575	0.07967874	0.6199127	0.7287595	0.5618357
	m	**ã**	**a**	**i**	**u**	
	0.918232	0.8417838	0.8510477	0.8460112	0.8596159	
f_0 min	**p**	**t**	**k**	**b**	**d**	**l**
	0.6958254	0.3859148	0.4709698	0.7850786	0.7009208	0.6633086
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.5288523	0.331708	0.242917	0.5961724	0.6946176	0.906724

	m	**ã**	**a**	**i**	**u**	
	0.9254605	0.8679158	0.7816426	0.8277855	0.8412675	
f_0 max	**p**	**t**	**k**	**b**	**d**	**l**
	0.4981731	0.3118255	0.3470546	0.8033394	0.7206845	0.5987123
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.4099499	0.170077	0.2021851	0.6384302	0.599929	0.5719882
	m	**ã**	**a**	**i**	**u**	
	0.8939246	0.7687469	0.8524131	0.8361274	0.8418792	
Durée	**p**	**t**	**k**	**b**	**d**	**l**
	0.07010801	0.06404275	0.03731679	0.278533	0.2910391	0.09352688
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.3424634	0.2946572	0.4316083	0.3455099	0.3597654	0.2336327
	m	**ã**	**a**	**i**	**u**	
	0.05910556	0.5357292	0.5533802	0.5252055	0.5718768	
MFCC1	**p**	**t**	**k**	**b**	**d**	**l**
	0.05774463	0.02257094	0.04827935	0.1516556	0.1637346	0.1879441
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.2122908	0.4058716	0.3514982	0.2511872	0.4723499	0.3063464
	m	**ã**	**a**	**i**	**u**	
	0.3779812	0.4770289	0.4015842	0.3448233	0.4900302	
MFCC2	**p**	**t**	**k**	**b**	**d**	**l**
	0.1001081	0.07249388	0.04668954	0.07302903	0.1213385	0.1849789
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.1566546	0.2836406	0.305663	0.2215188	0.1482659	0.2465727
	m	**ã**	**a**	**i**	**u**	
	0.4837475	0.503312	0.5162073	0.3644706	0.2830649	
MFCC3	**p**	**t**	**k**	**b**	**d**	**l**
	0.05833824	0.1130568	0.09323191	0.1657948	0.1523385	0.08783545
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.0442611	0.1799902	0.1590404	0.1811124	0.3532552	0.2416072
	m	**ã**	**a**	**i**	**u**	

	0.4378064	0.3607477	0.3204529	0.2427	0.2691307	
MFCC4	**p**	**t**	**k**	**b**	**d**	**l**
	0.03323571	0.09461923	0.08152469	0.1348117	0.1299538	0.0948641
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.2154873	0.05445722	0.2722623	0.1587467	0.4147271	0.2943845
	m	**ã**	**a**	**i**	**u**	
	0.469163	0.5433955	0.1928581	0.3156218	0.5437076	
MFCC5	**p**	**t**	**k**	**b**	**d**	**l**
	0.09594451	0.04432019	0.02919652	0.1888395	0.4102055	0.1323133
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.2144744	0.2302873	0.4570213	0.1658763	0.4259552	0.3038813
	m	**ã**	**a**	**i**	**u**	
	0.2570167	0.654421	0.2741703	0.6063977	0.3718761	
MFCC6	**p**	**t**	**k**	**b**	**d**	**l**
	0.02475961	0.07100519	0.07846521	0.1836734	0.3754667	0.3955127
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.2685208	0.06574844	0.1109335	0.2316089	0.4401823	0.1428141
	m	**ã**	**a**	**i**	**u**	
	0.6863397	0.4785744	0.3422865	0.6363452	0.4779155	
MFCC7	**p**	**t**	**k**	**b**	**d**	**l**
	0.03114324	0.02090095	0.03495946	0.1959078	0.2219873	0.08758104
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.146847	0.06181008	0.1570281	0.2478755	0.1446461	0.2465431
	m	**ã**	**a**	**i**	**u**	
	0.1519169	0.7656306	0.2697867	0.3015749	0.3118025	
MFCC8	**p**	**t**	**k**	**b**	**d**	**l**
	0.05238009	0.04571974	0.07757652	0.09176666	0.172777	0.2525524
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.2753352	0.1370277	0.1006121	0.1480397	0.3033157	0.09640006
	m	**ã**	**a**	**i**	**u**	
	0.352184	0.6248751	0.4567248	0.3614623	0.3393373	

MFCC9	p	t	k	b	d	l
	0.03702716	0.01581534	0.05452253	0.06150184	0.09396989	0.324794
	ʁ	f	ʃ	v	z	ʒ
	0.2505866	0.2055595	0.120744	0.2661665	0.1887784	0.3644979
	m	ã	a	i	u	
	0.3862543	0.7318528	0.5615895	0.5761204	0.5516208	
MFCC10	p	t	k	b	d	l
	0.009502097	0.0246827	0.02585671	0.05461027	0.1037067	0.355826
	ʁ	f	ʃ	v	z	ʒ
	0.2371461	0.03970265	0.2094006	0.2052216	0.1099683	0.4245751
	m	ã	a	i	u	
	0.2820602	0.7474476	0.2003355	0.4221046	0.4463509	
MFCC11	p	t	k	b	d	l
	0.02122141	0.06657549	0.07380154	0.09621035	0.1215723	0.1763522
	ʁ	f	ʃ	v	z	ʒ
	0.2438451	0.05442332	0.301246	0.1693574	0.1804004	0.1612395
	m	ã	a	i	u	
	0.3692795	0.3812573	0.4158018	0.3041218	0.4091935	
MFCC12	p	t	k	b	d	l
	0.0410171	0.03514164	0.03170445	0.06802352	0.1523876	0.3353601
	ʁ	f	ʃ	v	z	ʒ
	0.1704908	0.05085315	0.1784353	0.3439277	0.1785449	0.3588958
	m	ã	a	i	u	
	0.5140026	0.8141011	0.5512091	0.6808686	0.5465822	
MFCC13	p	t	k	b	d	l
	0.02152735	0.00524943	0.04444701	0.0724278	0.1991478	0.2843456
	ʁ	f	ʃ	v	z	ʒ
	0.2052132	0.06816755	0.2208929	0.2536504	0.2488504	0.4729752
	m	ã	a	i	u	
	0.2963192	0.3738985	0.4441558	0.2328614	0.3646414	

Valeurs de η2 - corpus Patatra, sous-ensemble femmes

MS1.1	**p**	**t**	**k**	**b**	**d**	**l**
	0.1095804	0.05479839	0.0584175	0.03223901	0.1908897	0.2649086
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.2216224	0.455985	0.3131209	0.1892991	0.4673114	0.2500065
	m	**ã**	**a**	**i**	**u**	
	0.3689429	0.4300227	0.3402198	0.4555141	0.275559	
MS1.2	**p**	**t**	**k**	**b**	**d**	**l**
	0.09268342	0.1518652	0.0558858	0.09335006	0.04089391	0.1249318
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.3404035	0.4145968	0.4666435	0.04960317	0.2601555	0.1425192
	m	**ã**	**a**	**i**	**u**	
	0.7312229	0.4679336	0.3400336	0.365963	0.1782387	
MS2.1	**p**	**t**	**k**	**b**	**d**	**l**
	0.04868297	0.0811828	0.1100887	0.07258723	0.1342933	0.2369966
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.135732	0.3221815	0.4174918	0.2362332	0.2920589	0.290896
	m	**ã**	**a**	**i**	**u**	
	0.2786328	0.502363	0.3264863	0.3691172	0.3038564	
MS2.2	**p**	**t**	**k**	**b**	**d**	**l**
	0.1283301	0.04078979	0.05555683	0.08644657	0.1249376	0.1918257
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.2734746	0.2793428	0.2865651	0.09023985	0.4368522	0.3016286
	m	**ã**	**a**	**i**	**u**	
	0.2077866	0.5050932	0.2173976	0.4043261	0.03136948	
MS3.1	**p**	**t**	**k**	**b**	**d**	**l**
	0.06063648	0.05226868	0.07629454	0.04995689	0.2741762	0.2894298
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.1541069	0.4385422	0.348385	0.1834535	0.4756243	0.2035069
	m	**ã**	**a**	**i**	**u**	
	0.2907324	0.5830485	0.31248	0.4612573	0.3546867	

MS3.2	p	t	k	b	d	l
	0.08584302	0.07018176	0.03220045	0.1712054	0.2861958	0.1467035
	ʁ	f	ʃ	v	z	ʒ
	0.0867111	0.2425652	0.1621102	0.2525468	0.3803268	0.3877457
	m	ã	a	i	u	
	0.1687318	0.3986226	0.2631793	0.3482958	0.2254634	
MS4.1	p	t	k	b	d	l
	0.03716344	0.0231781	0.05708316	0.04889647	0.240868	0.2685935
	ʁ	f	ʃ	v	z	ʒ
	0.0977964	0.2799073	0.5009589	0.15227	0.1724573	0.05098489
	m	ã	a	i	u	
	0.2383979	0.5746176	0.3257705	0.4555985	0.3189029	
MS4.2	p	t	k	b	d	l
	0.0539427	0.02977426	0.02828775	0.1550714	0.194843	0.1277484
	ʁ	f	ʃ	v	z	ʒ
	0.03047876	0.01327328	0.0275692	0.1656416	0.1869266	0.1822568
	m	ã	a	i	u	
	0.1208504	0.2212292	0.2299	0.2890973	0.1527224	
HNR mid	p	t	k	b	d	l
	0.03244404	0.09333245	0.05594496	0.04444685	0.05991638	0.1889024
	ʁ	f	ʃ	v	z	ʒ
	0.02277564	0.1682008	0.1570369	0.09403297	0.3420244	0.2729004
	m	ã	a	i	u	
	0.1921879	0.5978464	0.19155	0.1938132	0.251015	
HNR mean	p	t	k	b	d	l
	0.0902223	0.1246137	0.0211442	0.1996812	0.1913161	0.2705101
	ʁ	f	ʃ	v	z	ʒ
	0.1281843	0.1441788	0.2381844	0.3396223	0.2974972	0.2463313
	m	ã	a	i	u	
	0.1804738	0.7014843	0.2866785	0.2458535	0.3128552	
HNR sd	p	t	k	b	d	l

	0.08199993	0.08236734	0.03021823	0.2021901	0.1963831	0.1016446
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.1138575	0.1301649	0.1083201	0.0679984	0.1836166	0.08115369
	m	**ã**	**a**	**i**	**u**	
	0.3365048	0.2198066	0.1371288	0.0554775	0.04477248	
f_0	**p**	**t**	**k**	**b**	**d**	**l**
	0.5899159	0.9605682	0.6625759	0.5780411	0.5446381	0.2285184
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.09719361	0.2962235	0.2903748	0.2788168	0.4056234	0.4594146
	m	**ã**	**a**	**i**	**u**	
	0.7737252	0.4109191	0.6574448	0.628899	0.6458874	
f_0 min	**p**	**t**	**k**	**b**	**d**	**l**
	0.4440982	0.2824398	0.2851707	0.5130138	0.4074755	0.3163568
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.264318	0.1649491	0.231105	0.2985416	0.4031052	0.7717562
	m	**ã**	**a**	**i**	**u**	
	0.8013226	0.5366738	0.5135252	0.6165158	0.6380162	
f_0 max	**p**	**t**	**k**	**b**	**d**	**l**
	0.302137	0.1828478	0.193058	0.6065344	0.617819	0.270345
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.2006365	0.1143102	0.1197623	0.2979963	0.3319042	0.5196681
	m	**ã**	**a**	**i**	**u**	
	0.7990786	0.478555	0.652689	0.6104025	0.6105603	
Durée	**p**	**t**	**k**	**b**	**d**	**l**
	0.0954332	0.04265703	0.02525496	0.3496857	0.3496435	0.08814356
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.09266324	0.1182694	0.1131833	0.3265943	0.3847336	0.3273704
	m	**ã**	**a**	**i**	**u**	
	0.565282	0.5054301	0.1523353	0.07594614	0.07403385	
MFCC1	**p**	**t**	**k**	**b**	**d**	**l**
	0.04237538	0.02636172	0.01956841	0.1285722	0.2001545	0.09912806

	ʁ	f	ʃ	v	z	ʒ
	0.2342246	0.4238303	0.384703	0.1849797	0.4431378	0.1584822
	m	ã	a	i	u	
	0.2234652	0.4070945	0.3188139	0.2377902	0.2962004	
MFCC2	p	t	k	b	d	l
	0.06089773	0.1030299	0.07936817	0.08831373	0.155378	0.08163545
	ʁ	f	ʃ	v	z	ʒ
	0.08245639	0.3220911	0.324215	0.2958514	0.1467177	0.2648328
	m	ã	a	i	u	
	0.5312022	0.3587484	0.3356383	0.3748256	0.2199475	
MFCC3	p	t	k	b	d	l
	0.03183866	0.09634216	0.09929183	0.1938318	0.1484872	0.09356996
	ʁ	f	ʃ	v	z	ʒ
	0.0661402	0.1673772	0.07707952	0.1883452	0.2409637	0.1062849
	m	ã	a	i	u	
	0.3188565	0.423032	0.236485	0.1890243	0.2270798	
MFCC4	p	t	k	b	d	l
	0.03062405	0.1118588	0.03800472	0.13517	0.1354667	0.06746606
	ʁ	f	ʃ	v	z	ʒ
	0.1869818	0.07776795	0.2330736	0.1619148	0.1615936	0.2352594
	m	ã	a	i	u	
	0.5013949	0.3864444	0.1372161	0.2827382	0.2473858	
MFCC5	p	t	k	b	d	l
	0.1170239	0.05933898	0.02670021	0.2287437	0.4002637	0.08241261
	ʁ	f	ʃ	v	z	ʒ
	0.2989023	0.1273393	0.3994973	0.0594669	0.4079201	0.3138832
	m	ã	a	i	u	
	0.2697109	0.6288063	0.2524486	0.240953	0.2090429	
MFCC6	p	t	k	b	d	l
	0.02566126	0.06479819	0.06801263	0.1907197	0.2902389	0.06588823
	ʁ	f	ʃ	v	z	ʒ

	0.254895	0.05246269	0.1027871	0.1348192	0.2113058	0.05418572
	m	**ã**	**a**	**i**	**u**	
	0.6567182	0.5296947	0.4178505	0.3908475	0.370931	
MFCC7	**p**	**t**	**k**	**b**	**d**	**l**
	0.03372482	0.01173088	0.01538965	0.1837802	0.2501934	0.03941352
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.2062802	0.04087239	0.06580058	0.1863112	0.09999773	0.2341302
	m	**ã**	**a**	**i**	**u**	
	0.15945	0.7037712	0.1751456	0.3407982	0.4162684	
MFCC8	**p**	**t**	**k**	**b**	**d**	**l**
	0.07107229	0.06912723	0.08469444	0.1268249	0.2103828	0.05514634
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.1724895	0.1620707	0.07655118	0.07392807	0.3234239	0.1352028
	m	**ã**	**a**	**i**	**u**	
	0.2994446	0.5915613	0.3847008	0.2512802	0.2126189	
MFCC9	**p**	**t**	**k**	**b**	**d**	**l**
	0.01822069	0.01265667	0.01090526	0.06584249	0.06177559	0.2502663
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.07622315	0.2071282	0.2189048	0.1918219	0.2176829	0.336222
	m	**ã**	**a**	**i**	**u**	
	0.3932993	0.6858819	0.6066002	0.5559996	0.4654678	
MFCC10	**p**	**t**	**k**	**b**	**d**	**l**
	0.008102203	0.02883274	0.01246383	0.07451832	0.05076412	0.1109303
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.123954	0.05465126	0.1449761	0.1338837	0.09992607	0.352307
	m	**ã**	**a**	**i**	**u**	
	0.3520572	0.7728704	0.1287462	0.3234834	0.3594402	
MFCC11	**p**	**t**	**k**	**b**	**d**	**l**
	0.01824588	0.0324234	0.0622951	0.1190983	0.09274736	0.1218506
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.2266899	0.03899594	0.1618712	0.1746495	0.1427433	0.1950813

	m	**ã**	**a**	**i**	**u**	
	0.3980083	0.2488547	0.4926801	0.1220226	0.2147664	
MFCC12	**p**	**t**	**k**	**b**	**d**	**l**
	0.02715975	0.02544008	0.04048873	0.08724012	0.09028841	0.1852875
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.1010933	0.05210464	0.1866751	0.1724605	0.09512912	0.3024337
	m	**ã**	**a**	**i**	**u**	
	0.623429	0.8173767	0.4437408	0.2855233	0.3395069	
MFCC12	**p**	**t**	**k**	**b**	**d**	**l**
	0.00692620	0.00265827	0.05406882	0.04332013	0.09583974	0.1081347
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.03352198	0.06479118	0.07975568	0.1015442	0.08166268	0.1464063
	m	**ã**	**a**	**i**	**u**	
	0.1060899	0.3781342	0.2453811	0.09171899	0.4178188	

Valeurs de η2 - corpus Patatra, sous-ensemble hommes

MS1.1	**p**	**t**	**k**	**b**	**d**	**l**
	0.03415141	0.01106301	0.06603336	0.0566451	0.2916626	0.1572836
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.392797	0.3459676	0.4261849	0.2061914	0.118937	0.07645616
	m	**ã**	**a**	**i**	**u**	
	0.3169132	0.5763289	0.5650899	0.2381982	0.4208469	
MS1.2	**p**	**t**	**k**	**b**	**d**	**l**
	0.0772472	0.01290997	0.05096949	0.1616363	0.04978932	0.2101647
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.4077571	0.1816863	0.4500958	0.04403639	0.06272852	0.1078655
	m	**ã**	**a**	**i**	**u**	
	0.5337609	0.3088097	0.1591901	0.02934504	0.3876664	
MS2.1	**p**	**t**	**k**	**b**	**d**	**l**
	0.01214201	0.01700068	0.1062553	0.1370923	0.2939061	0.1136556
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**

	0.1926399	0.3388642	0.4441734	0.3098044	0.1780096	0.3414964
	m	**ã**	**a**	**i**	**u**	
	0.4527819	0.2911674	0.321338	0.2200059	0.3979948	
MS2.2	**p**	**t**	**k**	**b**	**d**	**l**
	0.04402196	0.00987585	0.02906862	0.0216628	0.0894477	0.03824148
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.4217594	0.3494676	0.2969888	0.1036367	0.142173	0.02238995
	m	**ã**	**a**	**i**	**u**	
	0.1357711	0.3623231	0.3168462	0.0904159	0.2046725	
MS3.1	**p**	**t**	**k**	**b**	**d**	**l**
	0.02662381	0.02194538	0.1129566	0.05608848	0.2784026	0.1601733
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.2330717	0.3560935	0.3426369	0.2380638	0.1814426	0.2772144
	m	**ã**	**a**	**i**	**u**	
	0.2744172	0.6287275	0.5719204	0.3469917	0.4961657	
MS3.2	**p**	**t**	**k**	**b**	**d**	**l**
	0.04586303	0.00738393	0.00563048	0.1445699	0.1433719	0.11528
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.3289041	0.01766512	0.2429244	0.03548703	0.1642327	0.1558446
	m	**ã**	**a**	**i**	**u**	
	0.359235	0.1230476	0.2009535	0.3037292	0.2611268	
MS4.1	**p**	**t**	**k**	**b**	**d**	**l**
	0.01776218	0.02682202	0.1372103	0.07038836	0.2475329	0.1491587
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.1134863	0.1766748	0.4596283	0.2019776	0.1046395	0.3776149
	m	**ã**	**a**	**i**	**u**	
	0.3188387	0.6096172	0.5100658	0.334956	0.4969495	
MS4.2	**p**	**t**	**k**	**b**	**d**	**l**
	0.04196865	0.00762827	0.01359306	0.11462	0.09957493	0.06634341
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.1883269	0.00851517	0.0712403	0.07148352	0.1052658	0.09726513

	m	**ã**	**a**	**i**	**u**	
	0.1868551	0.02149826	0.02160088	0.3610787	0.1227353	
HNR mid	**p**	**t**	**k**	**b**	**d**	**l**
	0.07075987	0.01887183	0.03658969	0.07879628	0.05983919	0.0612509
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.03748347	0.0169087	0.2593814	0.05591354	0.00895839	0.1545474
	m	**ã**	**a**	**i**	**u**	
	0.2878319	0.5418534	0.2756171	0.09437776	0.2469195	
HNR mean	**p**	**t**	**k**	**b**	**d**	**l**
	0.01671129	0.03508221	0.04054981	0.375046	0.4383918	0.2915662
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.07699859	0.08025994	0.334989	0.05958464	0.01335199	0.2568111
	m	**ã**	**a**	**i**	**u**	
	0.369021	0.6157119	0.3341873	0.228649	0.3258245	
HNR sd	**p**	**t**	**k**	**b**	**d**	**l**
	0.00856530	0.02147875	0.0384766	0.0536931	0.1212279	0.02464499
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.05425935	0.02013442	0.02519849	0.1240437	0.02622254	0.1275445
	m	**ã**	**a**	**i**	**u**	
	0.1881166	0.157498	0.01580436	0.03916634	0.06615506	
f_0	**p**	**t**	**k**	**b**	**d**	**l**
	/	/	/	0.7017567	0.6724374	0.7930849
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.06703981	0.3826038	0.00536771	0.4367111	0.648234	0.03887508
	m	**ã**	**a**	**i**	**u**	
	0.828192	0.7928084	0.8263107	0.722215	0.7714917	
f_0 min	**p**	**t**	**k**	**b**	**d**	**l**
	0.5254039	0.1069498	0.2443608	0.6912937	0.6802706	0.7244786
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.2872215	0.09533433	0.01806586	0.2776568	0.2964484	0.5975461
	m	**ã**	**a**	**i**	**u**	

	0.8414438	0.9031404	0.7941694	0.6708168	0.7075928	
f_0 max	**p**	**t**	**k**	**b**	**d**	**l**
	0.1277637	0.09140236	0.06133364	0.3722869	0.2623797	0.6781645
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.1531666	0.04442172	0.249328	0.7649697	0.2961004	0.04727712
	m	**ã**	**a**	**i**	**u**	
	0.5761597	0.6004974	0.8198063	0.7080959	0.7723605	
Durée	**p**	**t**	**k**	**b**	**d**	**l**
	0.01969547	0.08690129	0.05372178	0.08565596	0.04416696	0.1053428
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.5078938	0.4927958	0.7252124	0.3805276	0.2948442	0.05772328
	m	**ã**	**a**	**i**	**u**	
	0.1003984	0.6543834	0.05581236	0.01748065	0.03052145	
MFCC1	**p**	**t**	**k**	**b**	**d**	**l**
	0.06322795	0.01559706	0.0888463	0.1426906	0.05788547	0.08364392
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.1405101	0.3100314	0.2418268	0.2751588	0.06948928	0.03505183
	m	**ã**	**a**	**i**	**u**	
	0.4520194	0.2739796	0.2690622	0.2092764	0.4947857	
MFCC2	**p**	**t**	**k**	**b**	**d**	**l**
	0.08245262	0.01406467	0.00996435	0.02671935	0.00602725	0.2333482
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.193888	0.1865098	0.1613841	0.00824988	0.1039923	0.1745167
	m	**ã**	**a**	**i**	**u**	
	0.3411296	0.6402596	0.5615669	0.3360082	0.3919445	
MFCC3	**p**	**t**	**k**	**b**	**d**	**l**
	0.0267686	0.01450242	0.06625218	0.1084271	0.1593421	0.02810165
	ʁ	**f**	**ʃ**	**v**	**z**	**ʒ**
	0.00996464	0.2050966	0.1623034	0.1684593	0.4073988	0.2074058
	m	**ã**	**a**	**i**	**u**	
	0.5835957	0.1286144	0.2268868	0.2740894	0.3205757	

MFCC4	p	t	k	b	d	l
	0.0248642	0.00523544	0.07122819	0.109983	0.09165559	0.1003931
	ʁ	f	ʃ	v	z	ʒ
	0.1842859	0.01164255	0.2750534	0.04574126	0.1437453	0.3675772
	m	ã	a	i	u	
	0.2960579	0.4039575	0.2472303	0.01585415	0.5170304	
MFCC5	p	t	k	b	d	l
	0.05896668	0.01876204	0.01700081	0.06266043	0.237864	0.0251139
	ʁ	f	ʃ	v	z	ʒ
	0.01711277	0.2584944	0.5247706	0.317939	0.455523	0.1202713
	m	ã	a	i	u	
	0.1057003	0.5554133	0.2507881	0.4385249	0.3485172	
MFCC6	p	t	k	b	d	l
	0.02161651	0.07871346	0.04396408	0.01689647	0.2129021	0.2465438
	ʁ	f	ʃ	v	z	ʒ
	0.2288207	0.02761632	0.09839248	0.1567791	0.4701055	0.249846
	m	ã	a	i	u	
	0.550221	0.2289856	0.1013743	0.3530508	0.432838	
MFCC7	p	t	k	b	d	l
	0.02581548	0.00707702	0.03730513	0.1358572	0.1345929	0.03185396
	ʁ	f	ʃ	v	z	ʒ
	0.00639672	0.08263563	0.07830567	0.1497908	0.2131007	0.09712228
	m	ã	a	i	u	
	0.15945	0.5565484	0.1947558	0.2292096	0.08852065	
MFCC8	p	t	k	b	d	l
	0.01469622	0.00411622	0.06358844	0.05233572	0.07220739	0.3690389
	ʁ	f	ʃ	v	z	ʒ
	0.2951708	0.07902237	0.1411981	0.1523305	0.1191233	0.02722119
	m	ã	a	i	u	
	0.4146467	0.5719206	0.5505197	0.4811391	0.4647969	
MFCC9	p	t	k	b	d	l

	0.03055318	0.01601821	0.1067627	0.04522006	0.1119535	0.2095544
	ʁ	f	ʃ	v	z	ʒ
	0.1946178	0.169075	0.03788415	0.1253081	0.03798476	0.2243753
	m	ã	a	i	u	
	0.1673944	0.2657067	0.243069	0.252818	0.4831988	
MFCC10	**p**	**t**	**k**	**b**	**d**	**l**
	0.01240183	0.01703026	0.01833553	0.03337194	0.01022695	0.4528111
	ʁ	f	ʃ	v	z	ʒ
	0.2706208	0.00650750	0.03476662	0.06049327	0.07760146	0.2959065
	m	ã	a	i	u	
	0.1665966	0.6639621	0.30839	0.4908216	0.3775827	
MFCC11	**p**	**t**	**k**	**b**	**d**	**l**
	0.01956531	0.01305526	0.09022799	0.05601458	0.1432196	0.2350494
	ʁ	f	ʃ	v	z	ʒ
	0.05069675	0.07254313	0.160297	0.03106613	0.2027015	0.1076136
	m	ã	a	i	u	
	0.2929729	0.4950762	0.06149332	0.3715949	0.5516882	
MFCC12	**p**	**t**	**k**	**b**	**d**	**l**
	0.05930102	0.04988541	0.01295277	0.02598007	0.151233	0.0760096
	ʁ	f	ʃ	v	z	ʒ
	0.2745274	0.04930774	0.09947293	0.06229499	0.1339208	0.4224341
	m	ã	a	i	u	
	0.2076965	0.1152804	0.3991449	0.1441475	0.2701805	
MFCC13	**p**	**t**	**k**	**b**	**d**	**l**
	0.01184568	0.00785964	0.02237309	0.08620727	0.2583423	0.299287
	ʁ	f	ʃ	v	z	ʒ
	0.2315853	0.04515889	0.2287999	0.1268118	0.1705067	0.4673159
	m	ã	a	i	u	
	0.2113245	0.3248529	0.1984422	0.2912113	0.2522016	

Annexe 3 : tableaux de comparaisons des meilleurs êta carré par segment avec indices correspondants

Phonème	BREF (Kahn, 2011)		FABIOLE		PATATRA - HOMMES	
	Indice	η^2	Indice	η^2	Indice	η^2
p	cog	8	MFCC2	20	MFCC2	8
t	cog	13	MFCC2	15	durée	8
k	cog	10	MFCC2	12	MS4.1	13
b	cog	13	f_0	22	f_0	70
d	cog	13	f_0-min	22	f_0-min	68
l	cog	22	f_0	24	f_0	79
m	cog	39	MFCC7	33	f_0-min	84
ʁ	cog	10	MFCC3	12	durée	50
f	cog	19	MFCC3	14	durée	49
ʃ	cog	21	MFCC3	25	durée	72
v	cog	20	f_0-min	23	f_0-max	76
z	cog	24	f_0	18	f_0	64
ʒ	cog	27	/	/	f_0-min	59
ɑ̃	cog	41	MFCC7	41	f_0-min	90
ɛ̃	cog	53	MS1.1	37	/	/
ɔ̃	cog	45	MFCC7	38	/	/
a	f_4	45	MS1.1	30	f_0	82
i	cog	29	MFCC5	24	f_0	72
u	f_1	24	f_0-min	20	f_0-max	77

Phonème	BREF (Kahn, 2011)		PATATRA - FEMMES	
	Indice	η^2	Indice	η^2
p	cog	8	MS2.2	12
t	cog	8	MS1.2	15
k	cog	8	MS2.1	11
b	cog	11	f_0-max	60
d	cog	9	f_0-max	61
l	cog	17	f_0-min	31
m	cog	23	f_0-min	80
ʁ	cog	10	MS1.2	34
f	cog	14	MS1.1	45
ʃ	cog	18	MS4.1	50
v	cog	16	HNR-mean	33
z	cog	16	MS3.1	47
ʒ	cog	21	f_0-min	77
ã	cog	50	MFCC12	81
a	cog	51	f_0	65
i	f_2	30	f_0	62
u	f_3	26	f_0	64

Annexe 4 : tableau récapitulatif de pourcentages de contribution des variables dans l'ACP

Indice	FABIOLE				PATATRA			
	Brut		Normalisé		Brut		Normalisé	
	PC1	PC2	PC1	PC2	PC1	PC2	PC1	PC2
cog1	14.68	1.36	12.72	0.06	15.15	1.31	9.6	5.84
cog2	10.3	0.01	8.82	1.28	10.39	0.45	1.86	7.35
sd1	9.42	3.4	6.92	1.56	11.59	1.45	9.2	1.06
sd2	14.24	0.8	12.12	1.69	13.85	0.22	4.22	14.84
skw1	14.29	3.43	12.17	1.18	12.14	4.26	10.38	2.16
skw2	9.56	0.23	6.35	8.39	3.25	0.84	6.12	13.15
krt1	9.24	4.73	11.7	0.39	8.49	5.12	9.67	1.44
krt2	5.36	0.66	6.11	9.38	1.71	0.71	5.64	13.82
mpt hnr	0.59	4.24	0.91	8.01	3.36	3.04	6.24	0.01
mean hnr	3.05	8.87	0.44	18.83	7.51	3.45	5.91	3.46
sd hnr	0.06	0.08	1.67	13.13	0.2	0.81	0.9	4.61
f_0	5.27	16.32	6.55	10.63	8	8.33	10.59	2.46
min_f_0	1.83	28.75	7.09	9.57	0.28	34.7	10.32	3.16
max_f_0	1.94	26.85	6.36	10.4	0.47	35.22	9.01	3.52
durée	0.19	0.3	0.08	5.5	3.61	0.09	0.33	23.12

MFCC	FABIOLE		PATATRA	
	PC1	PC2	PC1	PC2
1	0	0	7.31	10.01
2	10.93	22.16	6.38	0.55
3	30.04	0.73	12.79	7.33
4	14.9	2.73	11.59	14.89
5	7.74	21.77	2.17	28.95
6	10.23	0.5	2.79	15.71
7	1.39	25.05	16.49	0.07
8	0.08	6.83	2.39	9.02
9	11.05	0.57	9.21	0.26
10	1.98	0.35	5.74	0.31
11	7.18	5.66	2.77	0.09
12	4.48	13.66	8.87	7.32
13	0	0	11.5	5.49

Bibliographie

Articles, documents et thèses

(AFCP, 2002) Association Francophone de la Communication Parlée (2002). Communiqué de l'AFCP du 3 Décembre 2002 concernant l'identification des individus par leur voix.

(Ahmad et al., 2015) Ahmad, J., Fiaz, M., Kwon, S., Sodanil, M., Vo, B., & Baik, S. W. (2015). Gender Identification using MFCC for Telephone Applications – A Comparative Study, IJCSEE Vol. 3, Issue 5.

(Ajili et al., 2016a) Ajili, M., Bonastre, J.-F., Ben Kheder, W., Rossato, S., & Kahn, J. (2016). Phonetic content impact on Forensic Voice Comparison (p. 210-217). IEEE.

(Ajili et al. 2016b) Ajili, M., Bonastre, J.-F., Kahn, J., Rossato, S., & Bernard, G. (2016). FABIOLE, a Speech Database For Forensic Speaker Comparison. LREC.

(Besse, 1992) Besse, P. C. (1992). PCA stability and choice of dimensionality, Statistics & Probability Letters 13, 405–410.

(Bidelman et Lee, 2015) Bidelman, G. M., & Lee, C.-C. (2015). Effects of language experience and stimulus context on the neural organization and categorical perception of speech. NeuroImage, 120, 191-200.

(Bidelman et al., 2013) Bidelman, G. M., Moreno, S., & Alain, C. (2013). Tracing the emergence of categorical speech perception in the human auditory system. NeuroImage, 79, 201-212.

(Bloomfield, 1933) Bloomfield, L., (1933). Language. Holt, Rinehart and Winston, New York.

(Blumstein et Stevens, 1981) Blumstein, S. E., & Stevens, K. N. (1981). Phonetic features and acoustic invariance in speech. Cognition, 10(1-3), 25-32.

(Boë, 2000) Boë, L.-J. (2000). Forensic voice identifcation in France. Speech Communication, 20.

(Boë et al., 2018) Boë, L.-J., Bimbot, F., Bonastre, J.-F., & Dupont, P. (2018). Des évaluations des systèmes de vérification du locuteur à la mise en cause des expertises vocales en identification juridique, 27.

(Boë et Bonastre, 2012) Boë, L.-J., & Bonastre, J.-F. (2012). Lidentification du locuteur : 20 ans de témoignage dans les cours de Justice. Le cas du LIPSADON laboratoire

indépendant de police scientifique. In Actes de la conférence conjointe JEP-TALN-RECITAL (p. 8).

(Bonastre, 2003) Bonastre, J.-F. (2003). Person Authentication by Voice: A Need for Caution (p. 4). Présenté à 8th European Conference on Speech Communication and Technology, EUROSPEECH 2003 - INTERSPEECH 2003, Geneva, Switzerland.

(Borràs-Comes et al., 2014) Borràs-Comes, J., Vanrell, M. del M., & Prieto, P. (2014). The role of pitch range in establishing intonational contrasts. Journal of the International Phonetic Association, 44(01), 1-20.

(Boujelbene et al., 2009) Boujelbene, S. Z., Mezghani, D. B. A., & Ellouze, N. (2009). Identification du Locuteur par Système Hybride GMM-SMO. In SETIT 2009 (p. 10). Tunisia.

(Boyer et al., 2001) Boyer, E., Adnet, C., Larzabal, P. & Petitdidier, M. (2001). Estimation conjointe de moments spectraux d'échos Doppler. Application au cyclone 'Georges', 4.

(Breiman et al., 1984) Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J., (1984). Classification and regression trees, Monterey, CA, Wadsworth & Brooks/Cole Advanced Books & Software.

(Brummer et Swart, 2014) Brummer, N., & Swart, A. (2014). A comparison of linear and non-linear calibrations for speaker recognition. Odyssey 2014: The Speaker and Language Recognition Workshop, 5.

(Cai et al., 2018) Cai, W., Doshi, A., & Valle, R. (2018). Attacking speaker recognition with Deep Generative Models, 5.

(Calliope, 1989) Calliope. (1989). La parole et son traitement automatique. Paris: Masson.

(Chadha et al., 2011) Chadha, A., Jyoti, D., & Roja, M. M. (2011). Text-Independent Speaker Recognition for Low SNR Environments with Encryption. International Journal of Computer Applications, 31(10), 8.

(Colombo et Bundy, 1983) Colombo, J., & Bundy, R. S. (1983). Infant response to auditory familiarity and novelty. Infant Behavior and Development, 6(2-3), 305-311.

(Davis et Mermelstein, 1980) Davis, S. B., & Mermelstein, P. (1980). Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. In IEEE Transactions on acoustics, speech and signal processign (Vol. ASSP-28).

(DeCasper et Fifer, 1980) DeCasper, A. J., & Fifier, W. P. (1980). Of human bounding: newborns prefer their mothers' voices. Science new series, 208(4448), 1174-1176.

(Dumpala et al., 2017) Dumpala, S. H., Panda, A., & Kopparapu, S. K. (2017). Improved I-vector-based Speaker Recognition for Utterances with Speaker Generated Non-speech sounds.

(Dunbar et al., 2017) Dunbar, E., Cao, X. N., Benjumea, J., Karadayi, J., Bernard, M., Besacier, L., Anguera, X., Dupoux, E. (2017). The zero resource speech challenge 2017 (p. 323-330). IEEE.

(Eskenazi, 1984) Eskenazi, M. (1984). Sur l'invariance vocalique en français. In Actes de 13ème édition des Journées d'Etude sur la Parole, JEP Bruxelles 1984 (p. 47).

(Ferragne et Pellegrino, 2010) Ferragne, E., & Pellegrino, F. (2010). Vowel systems and accent similarity in the British Isles: Exploiting multidimensional acoustic distances in phonetics. Journal of Phonetics, 38(4), 526–539.

(Floccia et al., 1997) Floccia, C., Christophe, A., & Bertoncini, J. (1997). High-amplitude sucking and newborns: The quest for underlying mechanisms. Journal of Experimental Child Psychology, 64, 175-198.

(Floccia et al., 2000) Floccia, C., Christophe, A., & Bertoncini, J. (2000). Unfamiliar voice discrimination for short stimuli in newborns, Development science Vol. 3, Issue 3.

(Fougeron et Keating, 1997) Fougeron, C., & Keating, P. A., (1997). Articulatory strengthening at edges of prosodic domains. The Journal of the Acoustical Society of America, 101(6), 3728-3740.

(GCP, 1990) Groupe de la Communication Parlée (1990). Motion adoptée à l'unanimité par le Bureau du GCP de la SFA, reconduite intégralement par le GFCP de la SFA en 1997 et par l'AFCP en 2002.

(Gendrot et al., 2012) Gendrot, C., Adda-Decker, M., & Schmid, C. (2012). Comparaison de parole journalistique et de parole spontanée : analyses de séquences entre pauses. Actes de la conférence conjointe JEP-TALN-RECITAL, 8.

(GFCP, 1999) Groupe Francophone de la Communication Parlée de la Société Française d'Acoustique (1999). Pétition pour l'arrêt des expertises vocales, tant qu'elles n'auront pas été validées scientifiquement.

(Gharsellaoui et al. 2018) Gharsellaoui, S., Selouani, S., A., Cichocki, W., Alotaibi, Y., & Dahmane, A., O. (2018). Application of the pairwise variability index of speech rhythm with particle swarm optimization to the classification of native and non-native accents. Computer Speech & Language, 48, 67-79.

(Greenberg, 2003) Greenberg, S. (2003). Pronunciation Variation is Key to Understanding Spoken Language, 4.

(Hermansky, 1990) Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis of speech. The Journal of the Acoustical Society of America, 87(4), 1738-1752.

(Hombert et Puech, 1984) Hombert, J. M., Puech, G. (1984). Variabilité et invariance : l'espace vocalique en swahili. In Actes de 13ème édition des Journées d'Etude sur la Parole, JEP Bruxelles 1984 (p. 97).

(Jolliffe, 1986) Jolliffe, I. T. (1986). Principal Component Analysis. Springer-Verlag.

(Jongman et al., 2000) Jongman, A., Wayland, R., & Wong, S. (2000). Acoustic characteristics of English fricatives. The Journal of the Acoustical Society of America, 108(3), 1252.

(Jousse, 2011) Jousse, V. (2011). Identification nommée du locuteur : exploitation conjointe du signal sonore et de sa transcription (Thèse de doctorat). Université du Maine.

(Kahn, 2011) Kahn, J. (2011). Parole de locuteur : performance et confiance en identification biométrique vocale (Thèse de doctorat). Université d'Avignon et pays de Vaucluse.

(Kamruzzaman et al., 2010) Kamruzzaman, S. M., Rezaul Karim, A. N. M., Saiful Islam, M., & Emdadul Haque, M. (2010). Speaker Identification using MFCC-Domain Support Vector Machine, 5.

(Keating, 1997) Keating, P. A. (1997). Word-level phonetic variation in large speech corpora.

(Khosravani et al., 2016) Khosravani, A., Glackin, C., Dugan, N., Chollet, G., & Cannings, N. (2016). The Intelligent Voice 2016 speaker recognition system. Intelligent Voice Limited.

(Levine et Hullett, 2002) Levine, T. R., & Hullett, C. R. (2002). Eta Squared, Partial Eta Squared, and Misreporting of Effect Size in Communication Research. Human Communication Research, 28(4), 14.

(Li et al., 2016) Li, L., Wang, D., Zhang, X., Zheng, T. F., & Jin, P. (2016). System combination for short utterance speaker recognition (p. 1-5). IEEE.

(Li et al., 2014) Li, L., Wang, D., Zhang, Z., & Zheng, T. F. (2014). Deep Speaker Vectors for Semi Text-independent Speaker Verification, 13(9), 5.

(Mahola et al., 2007) Mahola, U., Nelwamondo, F. V., & Marwala, T. (2007). HMM Speaker Identification Using Linear and Non-linear Merging Techniques, 6.

(Martin, 2008) Martin, P. (2008). Phonétique acoustique : introduction à l'analyse de la parole. Cursus. Paris : Armand Colin.

(Mehler et al., 1988) Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. Cognition, 29, 143-178.

(Nolan et Asu, 2009) Nolan, F., & Asu, E. L., (2009). The Pairwise Variability Index and Coexisting Rhythms in Language. Phonetica 66, 64–67.

(Ohala et Gilbert, 1979) Ohala J.J., Gilbert J.B., 1979, Listeners' Ability to identify Languages by their Prosody, in Problèmes de prosodie Vol.II, Expérimentations, modèles et fonctions, Studia Phonetica 18, 123-131.

(Oller, 2008) Oller, L. L. (2008). Analysis of Voice Signals for the Harmonics-to-Noise Crossover Frequency. KTH - School of Computer Science and Communication (CSC) Department of Speech, Music and Hearing, Barcelone.

(Osuna et al., 1997) Osuna, E., Freund, R., Girosi, F., (1997). Improved Training Algorithm for Support Vector Machines. Proc. IEEE NNSP '97.

(Pearson, 1901) Pearson, K. (1901). On Lines and Planes of Closest Fit to Systems of Points in Space. Philosophical Magazine, 2(11), 559–572.

(Pierrehumbert, 2003) Pierrehumbert, J. B. (2003). Phonetic Diversity, Statistical Learning, and Acquisition of Phonology. Language and Speech, 46(2-3), 115-154.

(Pierrehumbert et al., 2004) Pierrehumbert, J. B., Bent, T., Munson, B., Bradlow, A. R., & Bailey, J. M. (2004). The influence of sexual orientation on vowel production (L). The Journal of the Acoustical Society of America, 116(4), 1905-1908.

(Quinlan, 1986) Quinlan, J. R., (1986). Induction of Decision Trees. Kluwer Academic Publishers, Machine Learning 1, 81-106.

(Remez et al., 1997) Remez, R. E., Fellowes, J. M., & Rubin, P. (1997). Talker identification based on phonetic information. Journal of Experimental Psychology: Human Perception and Performance, 23(3), 16.

(Reynolds, 1995) D. A. Reynolds, 1995. Speaker identification and verification using gaussian mixture speaker models. Speech Communication 17(1-2), 91–108.

(Reynolds et al., 2000) Reynolds, D. A., Quatieri, T. F., & Dunn, R. B. (2000). Speaker Verification Using Adapted Gaussian Mixture Models. Digital Signal Processing, 10(1-3), 19-41.

(Richardson, 2011) Richardson, J. T. E. (2011). Eta squared and partial eta squared as measures of effect size in educational research. Educational Research Review, 6(2), 135-147.

(Richardson et al., 2015) Richardson, F., Reynolds, D., & Dehak, N. (2015). A unified Deep Neural Network for speaker and language recognition, 5.

(Rouvier et al., 2017) Rouvier, M., Bousquet, P.-M., Ajili, M., Ben Kheder, W., Matrouf, D., & Bonastre, J.-F. (2017). LIA system description for NIST SRE 2016, 5.

(Sadjadi et al., 2016a) Sadjadi, S. O., Ganapathy, S., & Pelecanos, J. (2016). The IBM 2016 Speaker Recognition System (p. 174-180).

(Sadjadi et al., 2016b) Sadjadi, S. O., Pelecanos, J., & Ganapathy, S. (2016). The IBM Speaker Recognition System: Recent Advances and Error Analysis.

(Safavi et al., 2012) Safavi, S., Najafian, M., Hanani, A., Russell, M., & Jan, P. (2012). Speaker Recognition for Children's Speech, 4.

(Sahidullah et Saha, 2013) Sahidullah, M., & Saha, G. (2013). A Novel Windowing Technique for Efficient Computation of MFCC for Speaker Recognition. IEEE Signal Processing Letters, 20(2).

(Sarkar et al., 2014) Sarkar, J., Saha, S., & Agrawal, S. (2014). An Efficient Use of Principal Component Analysis in Workload Characterization-A Study, AASRI Procedia, 8, 68-74.

(Schötz, 2002) Schötz, S. (2002). Paralinguistic Phonetics in NLP Models & Methods, NLP term paper 11.

(Serrurier et al., 2017) Serrurier, A., Badin, P., Boë, L.-J., Lamalle, L., & Neuschaefer-Rube, C. (2017). Inter-Speaker Variability: Speaker Normalisation and Quantitative Estimation of Articulatory Invariants in Speech Production for French (p. 2272-2276). ISCA.

(Shi et Werker, 2001) Shi, R., & Werker, J. F. (2001). Six-Month-Old Infants' Preference for Lexical Words. Psychological Science, 12(1), 70-75.

(Shi et Werker, 2003) Shi, R., & Werker, J. F. (2003). The Preference for lexical words basis of preference for lexical words in 6-month-old infants, Developmental Science 6(5).

(Shon, 2017) Shon, S. (2017). KU-ISPL speaker recognition systems, 5.

(Shon et al. 2017) Shon, S., Mun, S., & Ko, H. (2017). Recursive Whitening Transformation for Speaker Recognition on Language Mismatched Condition (p. 2869-2873). ISCA.

(Solé, 2003) Sole, M.-J. (2003). Is Variation Encoded in Phonology?, 15th ICPhS Barcelona.

(Société Française d'Acoustique, 1990) Société Française d'Acoustique. MOTION adoptée, le 7 septembre 1990, à Paris par le Bureau du Groupe Communication Parlée de la Société Française d'Acoustique.

(Spinu et Lilley, 2016) Spinu, L., & Lilley, J. (2016). A comparison of cepstral coefficients and spectral moments in the classification of Romanian fricatives. Journal of Phonetics, 57, 40-58.

(Toussaint, 2013) Toussaint, G. T. (2013). The Pairwise Variability Index as a Measure of Rhythm Complexity, 42.

(Trabelsi et Ayed, 2013) Trabelsi, I., & Ayed, D. B. (2013). A Multi Level Data Fusion Approach for Speaker Identification on Telephone Speech. Image Processing and Pattern Recognition, 6(2), 10.

(Vaissière, 1986) Vaissière, J. (1986). Variance and Invariance at the Word Level. J. S. Perkell & D. H. Klatt. Invariance and Variability in Speech Process, Lawrence Erlbaum Associates.

(Vaissière, 1995) Vaissière, J. (1995). Nasalité et Phonétique, Colloque sur le voile pathologique, Lyon, 12 mai 1995, Société Française de Phoniatrie et Groupe Francophone de la Communication Parlée.

(Vaissière, 2006) J. Vaissière, 2006. La phonétique. Paris : Presses Universitaires de France.

(Wang et Sun, 2017) Wang, Y., & Sun, W. (2017). Multi-speaker Recognition in Cocktail Party Problem. In CSPS (p. 8).

(Wieling et al., 2012) Wieling, M., Margaretha, E., & Nerbonne, J. (2012). Inducing a measure of phonetic similarity from pronunciation variation. Journal of Phonetics, 40(2).

(Wold et al., 1987) Wold, S., Esbensen, K., & Geladi, P. (1987). Principal Components Analysis, Chemometrics and Intelligent Laboratory Systems, 2 (1–3), 37-52.

(Yumoto et al., 1982) Yumoto, E., Wilbur, J. G., & Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. J. Acoust. Soc. Am., 71(6).

(Zhang et al., 2017a) Zhang, C., Bahmaninezhad, F., Ranjan, S., Yu, C., Shokouhi, N., & Hansen, J. H. L. (2017). UTD-CRSS Systems for 2016 NIST Speaker Recognition Evaluation (p. 1343-1347). ISCA.

(Zhang et al., 2017b) Zhang, M., Chen, Y., Li, L., & Wang, D. (2017). Speaker recognition with cough, laugh and « Wei » (p. 497-501). IEEE.

(Zhang, 2018) Zhang, X.-L. (2018). Linear Regression for Speaker Verification, 10.

(Zhao et al., 2017) Zhao, W., Gao, Y., & Singh, R. (2017). Speaker identification from the sound of the human breath, 5.

Sitographie et programmes informatiques

(AGNITiO, 2015) AGNITiO (2015). BATVOX Case Study - Cahuzac France. Consulté à l'adresse <http://www.agnitio-corp.com/resources/technology/case-studies/batvox-case-study-cahuzac-france>

(Apple Inc., 2017) Apple Inc. (2017). Hey Siri: An On-device DNN-powered Voice Trigger for Apple's Personal Assistant. Machine Learning Journal, Vol. 1, Issue 6. Consulté à l'adresse <https://machinelearning.apple.com/2017/10/01/hey-siri.html>

(Apple Inc., 2018) Apple Inc. (2018). Speech Synthesis Manager. Consulté à l'adresse <https://developer.apple.com/documentation/applicationservices/speech_synthesis_manager>

(Boersma et Weenink, version 6.0.37) Boersma, P., Weenink, D. Praat: doing phonetics by computer (Version 6.0.37). Amsterdam, Holland. Consulté à l'adresse <http://www.fon.hum.uva.nl/praat/>

(Brümmer et de Villiers, 2010) Brümmer, N., de Villiers, E. (2010) BOSARIS Toolkit. AGNITiO Research, South Africa. Consulté à l'adresse <https://sites.google.com/site/bosaristoolkit/>

(Free Software Foundation, 2007) Free Software Foundation, Inc. (2007). eSpeak: Speech Synthesizer. Consulté à l'adresse <http://espeak.sourceforge.net>

(Josse et Husson, 2008) Lê, S., Josse, J. & Husson, F. (2008). FactoMineR: An R Package for Multivariate Analysis. Journal of Statistical Software. 25(1). pp. 1-18. Consulté à l'adresse <http://factominer.free.fr/index.html>

(Microsoft Corporation, 2018) Microsoft Corporation. Speaker Recognition API. Consulté à l'adresse <https://azure.microsoft.com/en-gb/services/cognitive-services/speaker-recognition/>

(NIST-SRE 12) NIST. (2012). The NIST Year 2012 Speaker Recognition Evaluation Plan. Consulté à l'adresse <https://www.nist.gov/multimodal-information-group/speaker-recognition-evaluation-2012>

(NIST-SRE 16) NIST. (2016). NIST 2016 Speaker Recognition Evaluation Plan. Consulté à l'adresse <https://www.nist.gov/itl/iad/mig/speaker-recognition-evaluation-2016>

(Povey et al., 2011) Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., Vesely, K., (2011). The Kaldi Speech Recognition Toolkit. IEEE 2011 Workshop on

Automatic Speech Recognition and Understanding. Consulté à l'adresse <http://kaldi-asr.org/doc/index.html>

(R, version 3.4.4) R Foundation for Statistical Computing. R: A language and environment for statistical computing (Version 3.4.4). Vienna, Austria. Consulté à l'adresse <https://www.R-project.org/>

(Trouville, 2016) Trouville, R. (2016) PVI Calculator, avec des modifications mineures pour réadapter aux besoins de cette étude <roland.trouville@gmail.com>

(Wells, 1997) Wells, J.C. (1997). SAMPA computer readable phonetic alphabet. Consulté à l'adresse <http://www.phon.ucl.ac.uk/home/sampa/>