Détection précoce de la maladie de Parkinson par l'analyse de la voix et corrélations avec la neuroimagerie

Laetitia Jeancolas

université PARIS-SACLAY

TELECOM SudParis

NNT : 2019SACLL019

Thèse de doctorat

Détection précoce de la maladie de Parkinson par l'analyse de la voix et corrélations avec la neuroimagerie

Thèse de doctorat de l'Université Paris-Saclay
préparée à Télécom SudParis

École doctorale n°580 Sciences et technologies de l'information et de la communication (STIC)
Spécialité de doctorat : Traitement du signal et des images

Thèse présentée et soutenue à Evry, le 4 Décembre 2019, par

Laetitia Jeancolas

Composition du Jury :

Laurence Devillers Professeur, LIMSI-CNRS, Orsay, France	Présidente
Bjoern Schuller Professeur, Imperial College, Londres, UK	Rapporteur
Chafik Mokbel Professeur, Université de Balamand, Liban	Rapporteur
Serge Pinto Chargé de recherche CNRS, LPL, Aix en Provence, France	Examinateur
Stéphane Lehéricy Professeur, ICM (CENIR) Paris, France	Examinateur
Badr-Eddine Benkelfat Professeur, Télécom SudParis (SAMOVAR), Evry, France	Directeur de thèse
Habib Benali Professeur, Université Concordia, Montréal, Canada	Co-Directeur de thèse
Dijana Petrovska-Delacrétaz MdC, Télécom SudParis (SAMOVAR), Evry, France	Encadrante

Remerciements

Je tiens à remercier tout d'abord, les membres du jury, qui m'ont fait l'honneur de consacrer leur temps et leur expertise à la lecture et à l'évaluation de mon travail de thèse : Pr. Laurence Devillers, Pr. Bjoern Schuller, Pr. Chafik Mokbel, Dr. Serge Pinto et Pr. Stéphane Lehéricy. Je remercie chaleureusement mon encadrante Dr. Dijana Petrovska-Delacrétaz et mes deux directeurs de thèse, Pr. Habib Benali et Pr. Badr-Eddine Benkelfat, pour leur investissement, leurs conseils et les échanges scientifiques qu'on a pu avoir.

Je remercie l'école doctorale Sciences et technologies de l'information et de la communication (STIC), et notamment le responsable du pôle 1 Gilles Duc pour son écoute et ses conseils. Je remercie également l'Institut Mines-Télécom, la Fondation Mines-Télécom et l'Institut Carnot Télécom & Société numérique qui ont financé ma thèse par le biais du programme Futur et Ruptures.

Un grand merci à mes collègues de l'ICM et de l'hôpital la Pitié Salpêtrière, travaillant sur le protocole ICEBERG : Marie Vidailhet, Stéphane Lehéricy, Jean Christophe Corvol, Isabelle Arnulf, Graziella Mangone, Marie-Odile Habert, Nicolas Villain, Rahul Gaurav. Merci à eux de m'avoir intégrée à ce beau projet, et d'avoir rendu notre collaboration très agréable. Merci également à toute l'équipe du CIC, notamment Alizé, Christelle, Sandrine et Charlène, d'avoir pris le relai des acquisitions vocales. Merci aussi à la plateforme IRM du CENIR, notamment aux manipulateurs radio Ayoub, Mélanie et Stéphanie. Merci à Benoit Béranger pour son aide dans la mise en place de la séquence voix IRM, et à Mélanie Pelegrini pour ses conseils précieux concernant l'analyse de cette dernière (analyse que je n'ai pas pu intégrer à ce manuscrit, mais qui fera l'objet d'un prochain article).

Je remercie également mes collègues de l'équipe SAMOVAR, notamment Gérard Chollet pour son temps et ses conseils, et mes cobureaux Christian, Amine, Aymen, Nasri et Maxime pour leur gentillesse et tous les moments partagés. Un remerciement particulier à Amine pour les nombreux dépannages informatiques, et à Matheus qui a participé aux annotations. Un grand merci aussi à tous mes collègues qui ont accepté que je les enregistre pour augmenter ma base de données.

Enfin un immense merci à ma famille et mes amis pour leur soutien et leurs encouragements. Un grand merci en particulier à mon père pour son aide technique on ne peut plus précieuse dans la mise en place du serveur téléphonique et dans les annotations. Je tiens aussi à remercier ma mère pour la relecture de mon manuscrit, et pour son soutien moral, ainsi que ma grand mère. Un grand merci à Paul qui m'a soutenue du début à la fin, et à Tania qui a même fait le déplacement d'Afrique du Sud pour m'accompagner lors de la soutenance. Enfin merci à mes parents, oncles, tantes, cousines et parents d'amis qui ont accepté de participer aux enregistrements.

Pour finir je tiens à remercier tous les participants du protocole ICEBERG, sans qui cette recherche n'aurait pas pu avoir lieu.

Table des matières

Liste des principales abréviations

AMR	Adaptive Multi Rate
Acc	Accuracy
CMS	Cesptral Mean Subtraction
DatScan	123-I Ioflupane tomoscintigraphie d'émission monophotonique
DCL	Démence à Corps de Lévy
DDK	Diadococinésie
DET	Detection Error Tradeoff
DNN	Deep Neural Network
EER	Equal Error Rate
EM	Expectation Maximization
Fo	Fréquence fondamentale
GMM	Gaussian Mixture Model
GSM	Global System for Mobile Communications
IP	Internet Protocol
iRBD	idiopathic Rapid eye movement (REM) sleep Behavior Disorder
IRM	Imagerie par Résonance Magnétique
LDA	Linear Discriminant Analysis
LLH	log Likelihood
LSVT	Lee Silverman Voice Treatment
MAP	Maximum a Posteriori
MFCC	Mel Frequency Cepstral Coefficient
MP	Maladie de Parkinson
MPTP	Methyl-4-Phenyl-1,2,3,6-Tetrahydropyridine
MSA	Multiple system atrophy
NM	Neuromélanine
PCM	Pulse Code Modulation
PLDA	Probabilistic Linear Discriminant Analysis
PSP	Paralysie Supranucléaire Progressive
RSD	Relative Standard Deviation
RTC	Réseau téléphonique Commuté
SD	Standard Deviation
SIP	Session Initiation Protocole
SVM	Support Vector Machine
TEP	Tomographie par Emission de Positons
UBM	Universal Background Model
UKPDSBB	United Kingdom Parkinson's Disease Society Brain Bank
UPDRS	Unified Parkinson's Disease Rating Scale
VAD	Voice Activity Detection
VoIP	Voix sur IP
VQ	Vector Quantization

Avant propos

La Maladie de Parkinson (MP) est une maladie neurodégénérative, la 2^e plus courante après la maladie d'Alzheimer, elle se manifeste essentiellement par des troubles moteurs s'aggravant au cours du temps. Sa prévalence augmente avec l'âge : 1% des personnes âgées de plus de 60 ans sont touchées et jusqu'à 4% des plus de 80 ans [De Lau and Breteler, 2006]. Etant donné que notre espérance de vie est en constante progression, il en va de même avec le nombre de personnes atteintes de cette maladie. On connaît mal la cause de la maladie de Parkinson mais on sait qu'elle s'accompagne d'une raréfaction des neurones dopaminergiques dans la substance noire compacte (dans le mésencéphale), aboutissant à un défaut de libération de dopamine dans le striatum. A ce jour, le diagnostic repose principalement sur un examen clinique effectué par un médecin. Habituellement le diagnostic est posé si l'examinateur observe au moins deux des trois symptômes suivants : akinésie (lenteur d'initiation des mouvements), rigidité et tremblements au repos. Malheureusement ces symptômes moteurs ne se manifestent qu'après la perte de 50 à 60% des neurones dopaminergiques dans la substance noire [Haas et al., 2012] et de 60 à 80% de leur terminaisons striatales [Fearnley and Lees, 1991]. Un enjeu majeur de la recherche consiste donc à trouver des moyens de détecter plus précocement la maladie, afin de pouvoir à terme ralentir, voire stopper, sa progression dès le début.

Parmi les manifestations cliniques diverses de cette maladie, la modification de la voix des malades semble être un élément d'intérêt à plusieurs égards. Un grand nombre de publications existent sur l'étude de la voix dans la MP. Elles ont mis en évidence des perturbations appelées dysarthrie hypokinétique, qui se manifestent par une diminution de la prosodie (de l'intonation), des irrégularités dans la phonation et des difficultés d'articulation. Les précisions de classification rapportées sont comprises entre 70 à 99% pour les stades modérés à avancés de la maladie.

Quelques études se sont intéressées plus spécifiquement à la détection précoce de MP par la voix et ont rapporté des performances de détection allant de 70 à 90%, sur des bases de données constituées en moyenne de 20 MP et 20 sujets sains [Rusz et al., 2015, Orozco-Arroyave et al., 2016a, Novotný et al., 2014, Rusz et al., 2011b].

De plus certaines perturbations de la voix caractéristiques de le maladie de Parkinson semblent être déjà visibles plusieurs années avant le diagnostic clinique [Harel et al., 2004, Rusz et al., 2015]. Cependant il n'y a eu que très peu d'études qui se sont intéressées aux marqueurs pronostiques de cette maladie dans la voix au stade prodromique [Postuma, 2015].

D'autre part, certaines études ont exploré la possibilité d'un télédiagnostic de MP en utilisant des enregistrements vocaux réalisés avec des applications smartphones ou tablettes puis envoyés à un serveur distant pour analyse [Zhang et al., 2018, Benba et al., 2016b, Rusz et al., 2018, Vaiciukynas et al., 2017, Zhang, 2017, Sakar et al., 2017].

D'autres ont également exploré l'effet de la transmission de la voix via le réseau téléphonique sur la détection de MP (ou d'autres pathologies de la voix), en le simulant par une dégradation d'enregistrements de haute qualité [Wu et al., 2018, Tsanas et al., 2012a, Vásquez-Correa et al., 2017b, Fraile et al., 2009a].

Mais à notre connaissance aucune étude n'avait été publiée sur la détection de MP via des

enregistrements téléphoniques réels (issus du réseau téléphonique).

L'objectif de cette thèse est d'étudier les modifications de la voix aux stades débutant et prodromique de la maladie de Parkinson, et de développer des modèles de détection précoce, grâce à la constitution d'une grande base de données.

Nous nous intéresserons à l'analyse de la voix enregistrée en laboratoire et par téléphone, à différentes méthodes de classification, ainsi qu'à l'effet du genre sur la détection précoce de MP. Enfin, nous étudierons les éventuelles corrélations entre les paramètres vocaux et la neuroimagerie, afin de conclure si l'analyse de la voix peut aussi être utilisée pour suivre l'évolution de la maladie.

Le but étant à terme de pouvoir construire un outil de détection précoce et de suivi, peu couteux et accessible, utilisable par les médecins en cabinet ou par téléphone.

Chapitre 1

Introduction

1.1 La Maladie de Parkinson

La Maladie de Parkinson idiopathique (MP), décrite pour la première fois par James Parkinson en 1817, est une maladie neurodégénérative affectant le système nerveux central, qui se manifeste essentiellement par des troubles moteurs s'aggravant au cours du temps. C'est la 2^e maladie neurodégénérative la plus courante après la maladie d'Alzheimer. Sa prévalence augmente avec l'âge, étant donné que notre espérance de vie est en constante progression, il en va donc de même avec le nombre de personnes atteintes de cette maladie. 0n estime qu'elle touche actuellement 5 millions de personnes dans le monde et que 9 millions de personnes seront atteintes en 2030 [Dorsey et al., 2007]. Cette maladie atteint 1 % des personnes âgées de plus de 60 ans et jusqu'à 4% des plus de 80 ans [De Lau and Breteler, 2006]. En France, 170 000 personnes sont traités pour la maladie de Parkinson et environ 26 000 nouveaux cas sont diagnostiqués chaque année [Moisan, 2018]. Les hommes sont atteints environ 1,5 fois plus souvent que les femmes. L'âge moyen du diagnostic de la maladie est de 60 ans.

Les symptômes caractéristiques de la maladie de Parkinson sont l'akinésie (difficulté à initier les mouvements et lenteur d'exécution), la rigidité et les tremblements au repos. Des traitements existent mais diminuent seulement les symptômes, ils n'arrêtent pas ni ne ralentissent la progression de la maladie. Si le mécanisme de la maladie de Parkinson est connu, les causes demeurent inconnues mais résulteraient *a priori* d'une combinaison de facteurs environnementaux et génétiques prédisposants.

La maladie de Parkinson idiopathique est à distinguer des syndromes Parkinsoniens atypiques comme la démence à corps de Levy (DCL), l'atrophie multisystématisée (MSA) et la paralysie supranucléaire progressive (PSP), qui sont des pathologies associant souvent les symptômes de maladie de Parkinson avec d'autres troubles (comme des atteintes cognitives). Il faut également distinguer les syndromes parkinsoniens secondaires induits par des neuroleptiques ou exposition au Methyl-4-Phenyl-1,2,3,6-Tetrahydropyridine (MPTP). Il faut aussi différencier la MP d'autres maladies pouvant entraîner des tremblements, comme le tremblement essentiel, qui est un tremblement postural et non de repos.

1.1.1 Physiopathologie

La maladie de Parkinson est caractérisée par la raréfaction des neurones dopaminergiques dans la substance noire compacte (dans le mésencéphale), et une altération des connexions entre la substance noire compacte et le striatum, aboutissant à un défaut de libération de dopamine dans ce dernier. La dopamine est un neurotransmetteur présent dans deux grands systèmes : la voie mésocorticolimbique et le système nigrostrié (cf. Figure 1.1).

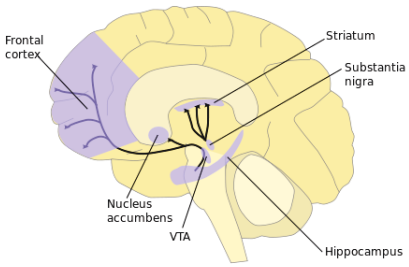


FIGURE 1.1 – Circuits dopaminergiques. Source : psychatricdrugs.com

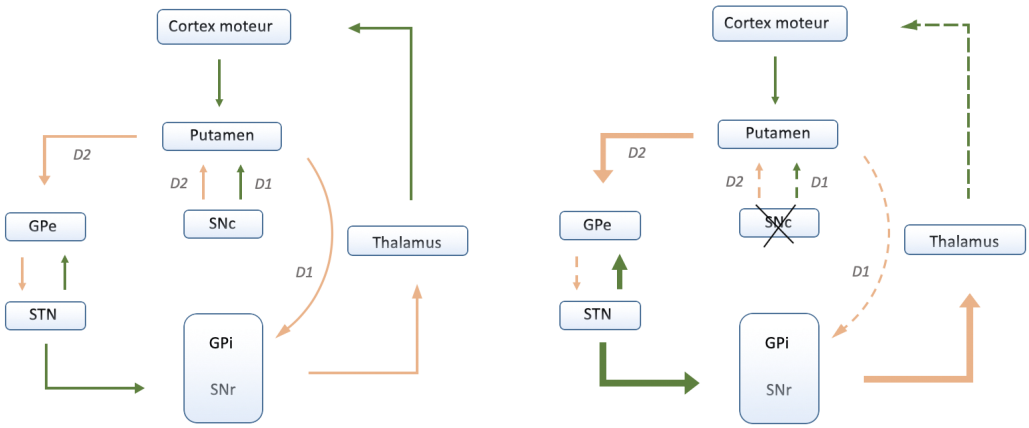
La voie mésocorticolimbique part de l'aire tegmentale ventrale et se projette au niveau du système limbique et du cortex frontal. Cette voie est responsable du circuit de la récompense, elle n'est *a priori* pas atteinte par la MP mais peut être impactée par les traitements médicamenteux donnés dans le cadre de MP, que ce soit des précurseurs de la dopamine (L-DOPA) ou des agonistes dopaminergiques.

Le système nigrostrié est lui composé de neurones dont la base est au niveau de la substance noire compacte et se projetant vers le striatum dorsal. Ce système est modulateur, entre autres, des aires corticales motrices. Il fait partie de deux grandes boucles cortico-sous-corticales impliquées dans le contrôle des fonctions motrices. Ces boucles font intervenir les ganglions de la base, le thalamus et le cortex (cf. Figure 1.2a). Les ganglions de la base, aussi appelés noyaux gris centraux, sont un ensemble de structures sous corticales composé de la substance noire compacte (SNc) et réticulée (SNr), du striatum, du noyau sous-thalamique (STN), et du globus pallidus interne (GPi) et externe (GPe). La première boucle contient la voie directe activatrice (D1), par désinhibition du thalamus. Dans cette boucle, le putamen (partie du striatum) est directement connecté au globus pallidus interne (GPi). Lorsque cette voie est activée, elle facilite les mouvements.

La deuxième boucle contient la voie indirecte inhibitrice du thalamus (D2), qui diminue les mouvements. Le putamen est lié au GPi par 2 intermédiaires (GPe puis STN). Lorsque la voie indirecte est activée, elle diminue les mouvements. Le système moteur est modulé par ces deux voies qui agissent sur le thalamus. Le thalamus projette ensuite sur le cortex moteur qui déclenche le mouvement au-delà d'un certain seuil. Ceci permet une gestion fine de l'activation de la voie motrice principale en assurant (voie directe) ou en inhibant (voie indirecte) le mouvement.

La substance noire compacte intervient dans ces boucles en facilitant D1 (par activation des neurones du putamen ayant des récepteurs dopaminergiques D1) et inhibant D2 (par inhibition de ceux qui ont des récepteurs D2). Dans la MP, l'altération de SNc conduit donc à une hypoactivation de D1 et à une hyperactivation de D2 (cf. Figure 1.2b). Ce modèle pathophysiologique expliquerait surtout les symptômes moteurs de bradykinesie [Rodriguez-Oroz et al., 2009]. La rigidité et les tremblements résulteraient d'un mécanisme plus complexe encore mal compris. Les ganglions de la base interviennent également dans des circuits associatifs et limbiques, qui comportent en plus une implication du noyau caudé (autre partie du striatum) et une projection dans respectivement le cortex préfrontal et le cortex cingulaire antérieur (cf. Figure 1.3) Dans MP la diminution dopaminergique au niveau du striatum cause ainsi également des dysfonctionnements cognitifs et thymiques.

Depuis 1998, la maladie de Parkinson est classée dans le groupe des synucléinopathies. Elle serait liée à des dépôts nommés corps de Lewy, qui se forment à l'intérieur des cellules nerveuse. Ce sont des agrégations d'une protéine naturellement présente dans le cerveau, l'alpha-synucléine, repliée de manière anormale. Récemment deux conformations de la protéine alpha-synucléine ont été identifiées, dont la forme fribrillaire qui serait impliquée dans la maladie de Parkinson [Pee-



(a) Modèle sain (b) Modèle MP

FIGURE 1.2 – Schémas de connectivité des ganglions de la base, circuit du contrôle moteur. En a) le modèle sain, en b) le modèle parkinsonien. Les flèches vertes (foncées) représentent les connexions excitatrices et les flèches orange (claires) les connexions inhibitrices. Dans le modèle parkinsonien, les connexions hypoactives sont représentées en pointillé et les connexions hyperactives en flèches épaisses. D1 est la voie directe désinhibitrice du thalamus (elle favorise le mouvement). D2 est la voie indirecte inhibitrice du thalamus (elle freine le mouvement). Dans la maladie de Parkinson D1 est sous-activée et D2 sur-activée.

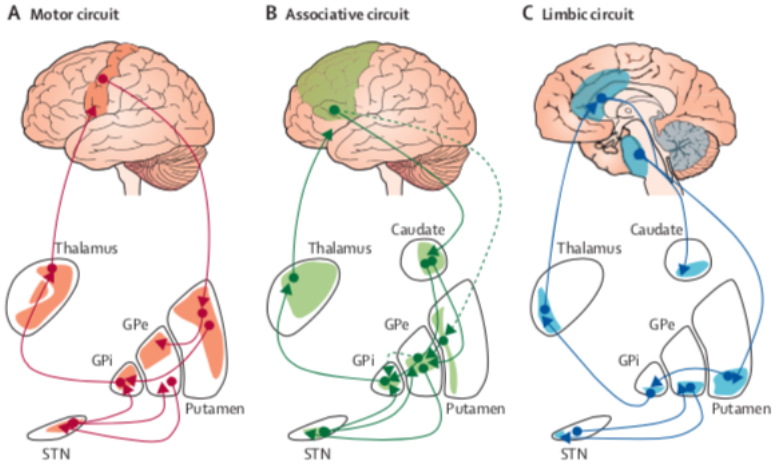


FIGURE 1.3 – Connectivité des ganglions de la base. Circuit moteur, associatif et limbique. Source : [Rodriguez-Oroz et al., 2009]

laerts et al., 2015]. Ces agrégats de protéines se propageraient à la manière des prions [Brundin et al., 2010].

Heiko Braak et son équipe proposent une description de la progression de la maladie en 6 niveaux [Braak et al., 2003]. Elle débuterait dans les noyaux moteurs dorsaux des nerfs vague et glossopharyngien et dans le bulbe olfactif. Puis se propageraient dans le mésencéphale (en touchant particulièrement la substance noire *pars compacta*). La progression toucherait ensuite le cortex, en commençant par la partie antéro-médiale temporale du mésocortex pour ensuite toucher le néocortex. La progression dans le néocortex commencerait par les zones associatives sensorielles et prémotrices pour atteindre finalement les aires primaires (sensorielles et motrices). L'apparition des symptômes moteurs coïnciderait avec la diminution de dopamine dans la partie

postérieure du putamen, qui correspond à la région motrice du striatum, soit au niveau 3 d'après Braak. Les niveaux 1 et 2 correspondraient à l'apparition de symptômes pré-moteurs (comme la diminution de l'odorat et le dérèglement du système digestif) et les niveaux 5 et 6 aux atteintes cognitives que l'on retrouve souvent en fin de MP (cf. Figure 1.4)

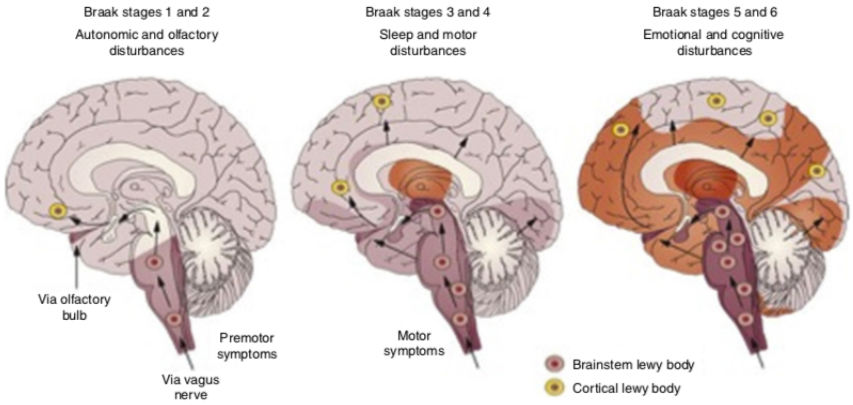


FIGURE 1.4 – Les différents niveaux de progression de la maladie de Parkinson selon Braak. Source : [Goldman and Holden, 2016]

1.1.2 Manifestations cliniques

1.1.2.1 Symptômes moteurs

Akinésie L'akinésie (difficulté à initier les mouvements) et la bradykinésie (lenteur d'exécution des mouvements) sont des symptômes moteurs très répandus dans la maladie de Parkinson. Les mouvements volontaires et involontaires (comme dans le maintien de l'équilibre, les expressions faciales, le clignement des yeux) sont altérés. L'akinésie rend compliqués le démarrage de la marche, les demi tours et les franchissements d'obstacles et peut entraîner une micrographie (l'écriture est plus petite et est exécutée plus lentement). Afin d'évaluer l'akinésie, le médecin peut demander aux patients de tapoter sur la table avec les doigts, de faire tourner les mains l'une autour de l'autre, d'exercer des mouvements circulaires avec un doigt, de taper le sol avec le talon et de marcher [Prescrire, 2015]. Le médecin va constater chez les patients MP un ralentissement et une diminution de l'amplitude des mouvements volontaires et une réduction des mouvements automatiques (ballant du bras, mimiques faciales).

Rigidité La rigidité est caractérisée par une augmentation involontaire du tonus musculaire. Elle touche les muscles fléchisseurs et extenseurs mais les muscles fléchisseurs des membres sont généralement affectés en premier, et de manière asymétrique. La rigidité se manifeste par une résistance au mouvement passif et peut être détectée par le test dit de roue dentée.

Tremblements Les tremblements débutent aux extrémités du corps et également de manière asymétrique. Ils se manifestent pendant le repos et sont accentués pendant le stress ou la concentration (par exemple pendant un exercice de calcul mental). Ils sont caractérisés par une contraction alternée des muscles agonistes et antagonistes à une fréquence de 4 à 6 Hz. Les tremblements sont le signe le plus souvent associé à la maladie de Parkinson. Pourtant 30 à 40% des personnes atteintes de MP n'ont pas de tremblements.

Instabilité posturale L'instabilité posturale est un symptôme moteur de MP qui arrive généralement tardivement dans MP et plus tôt dans les syndromes Parkinsoniens atypiques.

Il est provoqué par la perte des réflexes qui maintiennent une posture droite et est souvent source de chutes.

1.1.2.2 Symptômes non-moteurs

Les patients atteints de MP peuvent souffrir, en plus des symptômes moteurs, de nombreux symptômes non moteurs, affectant la cognition, l'humeur, les fonctions autonomes et le sommeil. La prévalence des symptômes non-moteurs dans la MP a été estimée à 98% d'après une étude multicentrique qui a examiné et interrogé 1072 patients avec MP [Barone et al., 2009]. Ces symptômes peuvent arriver aux différents stades de la maladie et s'améliorent généralement peu avec les traitements dopaminergiques, traduisant en plus de la dénervation nigrostriatale une atteinte plus diffuse et non dopaminergique [Lim et al., 2009].

Atteinte cognitive La maladie de Parkinson peut s'accompagner de troubles cognitifs allant d'une légère détérioration des fonctions exécutives (processus cognitif de haut niveau) à une démence (en fin de maladie). Les déficits cognitifs peuvent concerner l'attention, la mémoire (surtout l'étape de récupération), la planification, le traitement de l'information, la flexibilité mentale, la cognition sociale... Ils sont présents chez environ 30% des patients MP en début de maladie [Elgh et al., 2009].

Troubles psychiatriques Les symptômes psychiatriques rencontrés chez les patients débutant MP et non traités sont essentiellement : la dépression (37%), l'anxiété (17%) et l'apathie (27%) [Aarsland et al., 2009]. La dépression toucherait 70% des patients MP, tous stades confondus [Kulisevsky et al., 2008]. Elle serait un symptôme à part entière de MP et non juste une réactivité émotionnelle face à la maladie, pouvant arriver plusieurs années avant les symptômes moteurs [Santamaria et al., 1986]. L'anxiété peut y être associée et s'exprime sous forme d'attaques de panique ou de façon plus latente, et toucherait 69% des MP au cours de leur maladie d'après [Kulisevsky et al., 2008]. L'apathie (état émotionnel d'indifférence) concernerait quant à elle 48% des patients MP [Kulisevsky et al., 2008]. Elle peut être associée aux troubles anxio-dépressifs ou être présente seule [Pedersen et al., 2009]. D'autres troubles psychiatriques peuvent arriver en fin de maladie ou être induits par les traitements médicamenteux.

Douleur La douleur est présente chez 60% des MP [Barone et al., 2009]. Elle peut être due à l'hypertonie musculaire, prenant la forme de crampes ou douleurs musculaires continues. Elle peut aussi prendre la forme de perturbations sensitives (dysesthésie, sensations de brûlures, douleurs radiculaires) ou être secondaire à des problèmes articulatoires.

Dysautonomie La dysautonomie, (ou dystonie neurovégétative) correspond à un dérèglement global du système nerveux autonome. Dans la MP, plusieurs niveaux de ce système peuvent être touchés. Au niveau cardiovasculaire, on peut constater de l'hypotension orthostatique (chute brutale de la tension lors du passage de la position couchée à levée). Au niveau du système digestif, on peut rencontrer de l'hypersalivation, des difficultés de déglutition, de l'incontinence (urinaire et fécale), et de la constipation (signe pouvant précéder les symptômes moteurs de plusieurs années). Au niveau de la température, sa régulation peut également être atteinte, pouvant entraîner une sudation excessive.

Problèmes de sommeil : syndrome des jambes sans repos et RBD 90% des MP mentionnent des problèmes de sommeil, avec des difficultés à l'endormissement et des réveils fréquents, pouvant être dus, entre autres, au syndrome des jambes sans repos, ou à un trouble du comportement en sommeil paradoxal. Le syndrome des jambes sans repos, aussi appelé impatiences dans les jambes, est un trouble neurologique se caractérisant par un besoin irrépressible

de bouger les jambes, et se manifestant surtout le soir et la nuit. Il est souvent associé à la MP car il résulterait également d'un déficit de dopamine. Les troubles du comportement en sommeil paradoxal, ou *Rapid eye movement sleep Behaviour Disorder* en anglais (RBD) est une parasomnie caractérisée par une perte de l'atonie musculaire pendant la phase de sommeil paradoxal. Les patients atteints parlent, crient et bougent pendant qu'ils rêvent. Le diagnostic de RBD est posé après interrogatoire du patient, complété par un examen de vidéo-polysomnographie au cours duquel l'électromyographie révèle une tonicité musculaire élevée en phase de sommeil paradoxal, et durant lequel la vidéo peut montrer un comportement onirique anormal. 60 % des MP vont développer ce syndrome au cours de leur maladie [De Cock et al., 2007]. Les MP qui ont un RBD auraient une atteinte cognitive, d'après des tests neuropsychologiques, plus marquée que les MP qui n'ont pas ce trouble [Arnulf, 2012].

Vision des couleurs Un autre signe non moteur souvent rencontré dans la MP est une moins bonne discrimination des couleurs et une moins bonne sensibilité aux contrastes [Pieri et al., 2000].

1.1.2.3 Symptômes prodromiques

La phase prodromique est la période d'une maladie pendant laquelle des symptômes avant-coureurs, généralement bénins, annoncent la survenue de la phase principale de cette maladie. Ici nous considérons comme phase prodromique (ou préclinique) la phase qui commence au début du processus physiopathologique de MP et qui se termine au moment où le diagnostic clinique, tel qu'il est effectué actuellement (à partir des symptômes moteurs classiques), est possible. La difficulté résidant à savoir quand le processus biologique qui aboutira au diagnostic de MP commence. La perte dopaminergique commencerait 5 à 10 ans avant l'apparition des symptômes moteurs classiques [Meissner, 2012], et serait elle-même précédée de plusieurs années par des lésions au niveau du bulbe olfactif et de certains noyaux du tronc cérébral et de la moelle [Braak et al., 2003]. La durée totale de la phase préclinique serait ainsi de plusieurs décennies [Meissner, 2012]. Un ensemble de signes avant-coureurs de MP peuvent être rencontrés durant cette phase prodromique tels que des troubles de comportement en sommeil paradoxal, une diminution de l'odorat, de la constipation, une hypomimie faciale et des changements dans la voix.

iRBD *Rapid eye movement sleep Behaviour Disorder* (RBD) est un trouble qui touche environ 60 % des MP [De Cock et al., 2007] et qui se caractérise par une perte de l'atonie musculaire en sommeil paradoxal comme décrit dans le paragraphe 1.1.2.2. Tous les MP ne sont pas touchés par ce trouble mais à l'inverse quasiment toutes les personnes qui ont un RBD idiopathique (iRBD), donc sans MP associée, vont développer à terme un syndrome parkinsonien, le plus souvent la maladie de Parkinson, de temps en temps la démence à corps de Lévy et plus rarement une atrophie multisystématisée (MSA). Des études ont reporté un taux de conversion de 35% à 5 ans et 91% à 14 ans [Iranzo et al., 2014]. Ce syndrome est donc considéré comme un très bon marqueur précoce de synucléinopathie [Schenck et al., 2013]. Les iRBD qui développent MP le font à une moyenne d'âge de 70 ans, ce qui est 10 ans plus âgé que l'âge moyen d'apparition des symptômes moteurs chez l'ensemble des personnes atteintes de MP. De plus les MP qui ont commencé par un iRBD auraient plus d'atteintes cognitives que les autres MP, traduisant éventuellement un processus physiopathologique légèrement différent. Ainsi on peut considérer dans l'ensemble les iRBD comme étant au stade prodromique de MP, en gardant à l'esprit qu'ils ne sont peut être pas représentatifs de l'ensemble des patients au stade prodromique de MP.

Odorat L'hyposmie (diminution pathologique de l'odorat) touche environ 70 % des patients MP [Hawkes et al., 1997, Chen et al., 2015]. Elle est caractérisée par des difficultés à détecter les odeurs à les discriminer et à les identifier [Tissingh et al., 2001]. L'hyposmie est un des premiers

signes prodromiques à apparaître dans MP [Ponsen et al., 2004], ce qui est en accord avec l'accumulation de la protéine alpha-synucléine dans le bulbe olfactif au stade prodromique décrite par Braak [Braak et al., 2003]. Néanmoins l'hyposmie est un trouble assez courant touchant 25% des plus de 53 ans [Murphy, 2002]. Il ne peut donc est utilisé seul dans la détection précoce de MP.

Constipation La constipation est un symptôme gastro-intestinal courant dans la MP, pouvant résulter d'un dysfonctionnement du sphincter. Une grande étude longitudinale [Abbott et al., 2001] a montré que les hommes qui vont à la selle moins d'une fois par jour ont 2,7 fois plus de risque de développer MP que ceux qui y vont tous les jours et 4,1 fois plus de risques que ceux qui y vont 2 fois par jour. Le symptôme prodromique de constipation est cohérent avec l'accumulation précoce d'alpha synucléine décrite par Braak [Braak et al., 2003, Braak et al., 2006] au niveau des noyaux dorsaux du nerf vague et du système nerveux entérique. L'accumulation d'alpha synucléine dans le système nerveux entérique chez les MP a d'ailleurs été confirmée par des biopsies faites lors de coloscopies [Lebouvier et al., 2010]. Néanmoins, tout comme pour l'hyposmie, la constipation est un trouble courant [Peppas et al., 2008], ne permettant pas à elle seule de poser un diagnostic précoce ou non de MP.

Hypomimie faciale L'hypomimie (la réduction des mouvements du visage) est un symptôme concernant quasiment tous les patients atteints de la Maladie de Parkinson. Il est une manifestation de la bradykinesie et se manifeste par une diminution de l'expression faciale [Hoehn and Yahr, 1967] [Jankovic, 2003]. Ce symptôme est aussi présent chez les iRBD et précéderait le diagnostic de MP d'environ 7 ans [Postuma et al., 2012].

Changements vocaux Comme nous le détaillerons en partie 2.3, la voix, qui nécessite un contrôle moteur fin et une coordination précise entre les muscles laryngés (cordes vocales) et supralaryngés (langue, lèvres, mâchoire) subit beaucoup de modifications dans la MP (monotonie, difficultés d'articulation, modification de la fluence verbale, modifications du timbre...). Ces modifications arrivent tôt dans la maladie et peuvent précéder les symptômes moteurs classiques de plusieurs années. Des études rétrospectives et longitudinales ont montré que certaines perturbations vocales sont identifiables 5 à 7 ans avant le diagnostic de MP [Harel et al., 2004,Postuma et al., 2012].

1.1.3 Diagnostic

Actuellement, le diagnostic certain de la MP nécessite un examen anatomopathologique *post mortem* recherchant la présence de corps de Lévy dans le cerveau [Goldman and Holden, 2016]. En pratique un diagnostic probable peut être posé à la suite d'un examen clinique. Plusieurs critères de diagnostic existent et le plus courant est celui de United Kingdom Parkinson's Disease Society Brain Bank (UKPDSBB) [Hughes et al., 1992]. Ce critère diagnostique consiste en 3 étapes. La première sert à établir la présence d'un syndrome parkinsonien (défini par la présence de bradykinésie associée à une rigidité, des tremblements au repos, ou une instabilité posturale). La deuxième étape consiste à chercher des critères d'exclusion (qui seraient en faveur d'un syndrome parkinsonien atypique ou secondaire ou révéleraient des antécédents neurologiques). La troisième consiste à rechercher des éléments spécifiques à la MP (évolution lente et progressive des symptômes, asymétrie, bonne réponse à la L-DOPA..). Ce critère diagnostic permet de détecter MP avec 82% de réussite par rapport au diagnostic certain posé après autopsie [Hughes et al., 1992].

Les symptômes moteurs sur lesquels se fonde ce critère diagnostique ne se manifestent qu'après la perte de 50 à 60% des neurones dopaminergiques dans la substance noire [Haas

et al., 2012] et de 60 à 80% de leur terminaisons striatales [Fearnley and Lees, 1991], [Hornykiewicz, 1998]. Ce qui implique un diagnostic plutôt tardif dans la progression pathophysiologique de la maladie. Un enjeu majeur de la recherche médicale consiste donc à trouver des moyens de détecter plus précocement la maladie, afin de pouvoir à terme ralentir, voire stopper, sa progression dès le début.

1.1.4 Quantification de la progression

Une fois le diagnostic posé, afin de quantifier la progression de la maladie et les effets des traitements, plusieurs échelles sont habituellement utilisées. La première, l'échelle de Hoehn et Yahr date de 1967 [Hoehn and Yahr, 1967]. Elle divise la progression de la maladie en 5 niveaux, allant de la présence de signes unilatéraux entrainant un handicap fonctionnel minime ou nul (stade 1) jusqu'au confinement à la chaise roulante ou au lit avec perte d'autonomie (stade 5).

Depuis une échelle plus complète, l'*Unified Parkinson's Disease Rating Scale* (UPDRS) l'a remplacée dans la majorité des études cliniques sur la MP. Elle est composée de 6 sections qui reposent sur des interrogatoires ou des observations cliniques. La partie III, qui comporte une évaluation motrice est particulièrement utilisée pour évaluer la progression de la maladie.

En 2007 la Movement Disorder Society (MDS) a publié une révision de cette échelle [Goetz et al., 2007] nommée MDS-UPDRS. Cette nouvelle échelle est plus sensible, intègre les symptômes non moteurs et une notation des items homogénéisée. Cette échelle est composée de 4 parties qui contiennent plusieurs items, chaque réponse étant notée de 0 (=normal) à 4 (=sévère).

- La partie 1 relate les expériences non motrices de la vie quotidienne. Elle est complétée en partie par l'investigateur (1A) à partir de l'interrogatoire avec le patient et en partie (1B) par le patient, aidé éventuellement d'un aidant.
- La partie 2 traite des expériences motrices de la vie quotidienne. Elle consiste en un questionnaire auto-administré par le patient comme dans la partie 1B.
- La partie 3 est un examen moteur. Elle consiste en une série de tests moteurs à partir desquels l'investigateur va évaluer la parole, l'expression faciale, la rigidité, les mouvements des doigts, de la main, des orteils, des jambes, le lever du fauteuil, la marche, la posture et les tremblements. Pour les patients déjà sous traitement, cette partie est habituellement faite en ON (dans les 3 heures qui suivent la prise du traitement) et en OFF (plus de 12h depuis la dernière prise médicamenteuse) pour voir l'effet du traitement.
- La partie 4 vise à évaluer les complications motrices. L'investigateur à partir de l'interrogatoire et de son observation évalue les dyskinésies dues aux traitements et les fluctuations motrices incluant la dystonie en état OFF.

1.1.5 Examens complémentaires

Des examens de neuroimagerie peuvent être effectués en complément de l'analyse clinique pour aider au diagnostic ou au suivi de la MP.

Le scanner cérébral est normal dans la MP, mais permet d'éliminer des pathologies d'allure pseudo-parkinsonienne en détectant d'éventuelles lésions d'évolution progressive [Viallet et al., 2010].

L'Imagerie par Résonance Magnétique (IRM) standard est également normale dans la MP idiopathique mais permet d'éliminer d'autres syndromes parkinsoniens, comme la MSA et la PLP.

L'électroencéphalographie (EEG) peut également contribuer au diagnostic différentiel entre la MP et la DCL. L'électromyographie (EMG) peut être utilisé, quant à lui, pour un diagnostic

différentiel avec la MSA.

La Tomographie par Emission de Positon (TEP) permet de détecter la MP tôt dans le développement de la maladie. Elle montre une réduction asymétrique de la concentration du traceur (souvent le fluor 18) au niveau du putamen, cf. Figure 1.5. Cette technique permet de détecter avec une bonne précision la présence ou non d'un syndrome parkinsonien, mais a un faible pouvoir discriminant entre ces syndromes, rendant difficile le diagnostic différentiel entre la MP, la MSA et la PLP (d'après [Tolosa et al., 2006]). De plus le coût élevé de cet examen et sa faible disponibilité ne permet pas une utilisation en clinique, son utilisation est restreinte au cadre de la recherche.

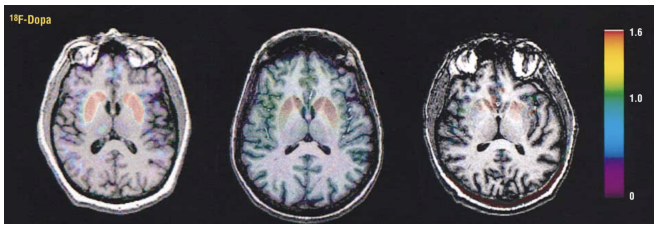


FIGURE 1.5 – Images TEP au niveau du striatum d'un sujet sain (à gauche), d'un patient MP débutant non traité (au milieu) et d'un patient MP avancé (à droite). Source : [Ribeiro et al., 2002]

La tomographie par émission monophotonique, ou SPECT (single photon emission computed tomography), est une technique de scintigraphie qui, utilisée avec le traceur Ioflupane(I-123) marqué à l'iode 123, permet d'étudier la fonction dopaminergique pré-synaptique. Cet examen, aussi appelé DatScan, met en évidence une diminution des transporteur dopaminergiques au niveau du striatum dans la MP (cf. Figure 9.1) . Cet examen est fiable mais ne permet pas de différencier la MP des syndromes parkinsoniens atypiques, de même son coût élevé fait qu'il est peu utilisé en clinique.

1.1.6 Traitements

Les traitements médicamenteux donnés pour réduire les symptômes moteurs de MP cherchent à pallier la diminution de dopamine au niveau du striatum. Deux types de médicaments sont habituellement donnés : les précurseurs de dopamine (L-DOPA) et des agonistes dopaminergiques. La L-DOPA est l'isomère lévogyre du 3,4-dihydroxyphénylalanine (abrévié DOPA). Elle a la particularité de traverser la barrière hémato-encéphalique et sa décarboxylation par la DOPA-décarboxylase produit la dopamine. C'est le traitement le plus actif et le plus donné dans la maladie de Parkinson. Sa durée de demi-vie est de 1.5 à 3h. Elle a comme effet secondaire d'entraîner des dyskinésies chez 40% des gens après quelques années de traitement [Ahlskog and Muenter, 2001] et des troubles compulsifs chez 10% des personnes traitées [Weintraub et al., 2010]. La deuxième grande classe de médicament prescrite pour la MP est celle des agonistes dopaminergiques, qui imitent la dopamine en se plaçant directement sur les récepteurs postsynaptiques de la voie nigrostriée. Cette classe de médicament est moins efficace que la L-DOPA mais peut être prescrite chez des sujets jeunes afin d'éviter une apparition trop précoce des dyskinésies induites par la L-DOPA. Ils ont une durée de demi-vie plus longue que la L-DOPA.

Pour les patients pour qui les traitements médicamenteux ne seraient pas assez efficaces, un traitement chirurgical est possible : la stimulation cérébrale profonde. Des électrodes sont implantées dans des régions précises du cerveau (noyau sous-thalamique, ou globus pallidus) et

y délivrent des courants électriques de faible intensité. Le mode d'action n'est pas encore bien compris, mais son effet sur les différents réseaux neuronaux fonctionnels a été mis en évidence récemment [Kibleur, 2016], et son efficacité prouvée sur des patients résistants au traitement habituel.

Les traitements médicamenteux et chirurgicaux réduisent les symptômes moteurs associés à la MP mais ne ralentissent pas l'évolution de la maladie. L'intérêt de la recherche sur le diagnostic précoce de MP est de pouvoir, dans un futur proche, tester de nouveaux traitements (dont le but serait d'arrêter la progression de la maladie) au stade prodromique de MP, quand le cerveau n'a pas encore trop de lésions irréversibles. Et une fois qu'un tel traitement sera découvert, de le donner le plus tôt possible aux patients atteints de MP afin de stopper dès le début la progression de la maladie.

1.2 Protocole ICEBERG

L'étude ICEBERG, dans laquelle s'insère ma thèse, est un protocole longitudinal monocentrique INSERM, qui a commencé en 2014, sur la recherche de marqueurs prédictifs de la maladie de Parkinson. Environ 300 sujets ont été recrutés, dont des MP idiopathiques débutants (moins de 4 ans d'évolution), des sujets iRBD (donc *a priori* en phase prodromique de MP) et des sujets sains. Les sujets viennent une fois par an à l'hôpital de la Pitié Salpêtrière et ce pendant 4 ans. Ils effectuent une batterie d'examens (prises de sang, DatScan, IRM, tests moteurs, bilan neuropsychologique, test d'odorat...) au cours de leurs visites. Ce protocole a reçu les autorisations des comités d'éthiques et de l'ANSM.

Depuis 2016, je propose aux sujets ICEBERG en plus des autres tests un enregistrement de la voix lors de leur visite annuelle, des enregistrements vocaux mensuels par téléphone, et une séquence d'IRM fonctionnelle, ajoutée à leur IRM, pendant laquelle ils effectuent une tâche vocale.

Pour pallier le nombre restreint de sujets sains inclus au début de ma thèse, j'ai aussi constitué une base de données supplémentaires en enregistrant des sujets sains de mon entourage (cf. section 4).

1.3 Contexte et objectifs de la thèse

A ce jour, le diagnostic de la MP repose principalement sur un examen clinique effectué par un neurologue. Habituellement le diagnostic est posé s'il observe au moins deux des trois symptômes suivants : akinésie (lenteur d'initiation des mouvements), rigidité et tremblements au repos. Malheureusement ces symptômes moteurs ne se manifestent qu'après la perte de 50 à 60% des neurones dopaminergiques dans la substance noire [Haas et al., 2012] et de 60 à 80% de leur terminaisons striatales [Fearnley and Lees, 1991]. Un enjeu majeur de la recherche consiste donc à trouver des moyens de détecter plus précocement la maladie, afin de pouvoir à terme ralentir, voire stopper, sa progression dès le début.

Parmi les manifestations cliniques diverses de cette maladie, la modification de la voix des malades semble être un élément d'intérêt à plusieurs égards. Un grand nombre de publications existent sur l'étude de la voix dans la maladie de Parkinson. Elles ont mis en évidence un ensemble de perturbations vocales, résumées par le terme de dysarthrie hypokinétique, introduit par [Darley Frederic L. et al., 1969]. Cette notion regroupe principalement des altérations prosodiques (diminution de l'intonation), articulatoires (imprécision des consonnes et des voyelles), et phonatoires (timbre rauque et soufflé). Les précisions de classification (*accuracy*) rapportées dans la littérature sont comprises entre 70 à 99% pour les stades modérés à avancés de la maladie.

Quelques études se sont intéressées plus spécifiquement à la détection précoce de MP par la voix et ont rapporté des performances de détection allant de 70 à 90%, sur des bases de données constituées en moyenne de 20 MP et 20 sujets sains [Rusz et al., 2015, Orozco-Arroyave et al., 2016a, Novotný et al., 2014, Rusz et al., 2011b].

De plus certaines perturbations de la voix caractéristiques de le maladie de Parkinson semblent être déjà visibles plusieurs années avant le diagnostic clinique [Harel et al., 2004, Postuma et al., 2012, Rusz et al., 2015]. Cependant il n'y a eu que très peu d'études qui se sont intéressées aux marqueurs pronostiques de cette maladie dans la voix au stade prodromique [Postuma, 2015].

D'autre part, certaines études ont exploré la possibilité d'un télédiagnostic de MP en utilisant des enregistrements vocaux réalisés avec des applications smartphones ou tablettes puis envoyés à un serveur distant pour analyse [Zhang et al., 2018, Benba et al., 2016b, Rusz et al., 2018, Vaiciukynas et al., 2017, Zhang, 2017, Sakar et al., 2017].

D'autres ont également exploré l'effet de la transmission de la voix via le réseau téléphonique sur la détection de MP (ou d'autres pathologies de la voix), en le simulant par une dégradation d'enregistrements de haute qualité [Wu et al., 2018, Tsanas et al., 2012a, Vásquez-Correa et al., 2017b, Fraile et al., 2009a].

Mais à notre connaissance aucune étude n'avait été publiée sur la détection de MP via des enregistrements téléphoniques réels (issus du réseau téléphonique).

L'objectif de cette thèse est d'étudier les modifications de la voix aux stades débutant et prodromique de la maladie de Parkinson, à partir d'une grande base de données. Nous nous intéresserons à l'analyse de la voix enregistrée en laboratoire et par téléphone. Le but étant à terme de pouvoir construire un outil de détection précoce peu couteux et accessible, utilisable par les médecins en cabinet, et une aide au diagnostic précoce par téléphone. Les objectifs principaux sont les suivants :

- Etude des caractéristiques de la voix aux stades débutant et prodromique de MP.
- Prédiction de MP, entièrement automatique, à partir de ces caractéristiques.
- Suivi de l'évolution de MP par l'analyse de la voix, et corrélations avec la neuroimagerie.

Pour cela nous avons constitué une base de données voix de plus de 200 locuteurs français, comprenant des sujets MP débutants (dont le diagnostic remontait à moins de 4 ans), des sujets sains et des sujets iRBD, considérés au stade prodromique de la maladie de Parkinson. Les sujets ont été enregistrés pendant une quinzaine de minutes avec un microphone professionnel, et en simultané avec le microphone interne d'un ordinateur. Ils ont également effectué une fois par mois des enregistrements vocaux, en appelant un serveur vocal interactif, à partir de leur propre téléphone. Au cours de ces enregistrements les sujets ont effectué différentes tâches vocales, comme des voyelles soutenues, des répétitions de phrases, de la lecture, des répétitions rapides de syllabes (DDK), des répétitions lentes de syllabes à un rythme imposé, et un monologue au cours duquel ils devaient raconter leur journée.

Nous avons analysé ces enregistrements vocaux par le biais de trois méthodes d'analyses différentes, couvrant différentes échelles de temps et différents domaines de la voix. Les deux premières méthodes sont inspirées de méthodes utilisées en reconnaissance du locuteur. Elles utilisent toutes les deux des paramètres court-terme caractérisant l'enveloppe spectrale, donc plutôt liés à l'articulation. La troisième méthode utilise des paramètres globaux reflétant d'autres domaines de la voix, comme la prosodie, la phonation, la fluence verbale, et la capacité à suivre un rythme imposé. Enfin, nous avons effectué une fusion de ces trois méthodes. Toutes les analyses ont été faites en traitant séparément les hommes des femmes, afin de ne pas rajouter la variabi-

lité due au genre, et afin d'évaluer d'éventuelles différences, selon le genre, dans les changements vocaux dus à MP.

L'originalité de cette thèse par rapport à l'état de l'art est :

- l'utilisation d'une grande base de données, composée de plus de 200 sujets ;
- l'analyse de la voix sur des sujets MP débutants traités, donc plus difficiles à détecter ;
- l'utilisation de la langue française pour détecter la MP ;
- l'effet du genre sur la détection précoce de MP ;
- l'influence du microphone utilisé sur les performances de détection de MP ;
- l'analyse de la voix MP via des enregistrements téléphoniques réels ;
- la couverture de tous les domaines de la voix (prosodie, articulation, phonation, fluence, et capacité à suivre un rythme imposé) ;
- l'utilisation de méthodes de classification encore jamais utilisées dans la détection de MP ;
- la mise en évidence de corrélations entre les paramètres vocaux et les changements dans le système dopaminergique quantifiés par la neurorimagerie.

L'organisation de ce manuscrit suit la logique suivante. Nous commençons, chapitre 2, par un rappel du fonctionnement de la voix et de son analyse, suivi d'un état de l'art sur ses modifications dans la MP. Ensuite chapitre 3, nous présentons un état de l'art des techniques de classifications utilisées, et les performances obtenues, pour la détection de MP via des analyses de la voix. Chapitre 4 nous décrivons les bases de données que nous avons constituées. Ensuite nous détaillons nos analyses par type de méthode utilisée. Les chapitres 5 et 6 s'appuient sur des méthodes inspirées de la reconnaissance du locuteur. Elles utilisent toutes les deux des paramètres cepstraux, les *Mel Frequency Cepstral Coefficients* (MFCC), caractérisant l'enveloppe spectrale, donc plutôt liés à l'articulation. Pour la méthode MFCC-GMM (*Gaussian Mixture Model*), la classification s'opère à l'échelle de la trame (fenêtre de 20ms), alors que pour la méthode des x-vecteurs, la classification se fait au niveau du segment (3s). La troisième méthode, détaillée dans le chapitre 7, utilise des paramètres dit globaux, calculés à l'échelle des tâches, et reflétant d'autres domaines de la voix, comme la prosodie, la phonation, la fluence verbale, et la capacité à suivre un rythme imposé. S'ensuit un chapitre sur la fusion de ces trois méthodes d'analyse, avec la présentation des résultats finaux de classification. Enfin dans le chapitre 9 nous présentons une analyse de corrélations entre les paramètres vocaux et l'évolution de la maladie, quantifiée par la neuroimagerie et des paramètres cliniques. Nous terminons par un chapitre de conclusion générale et de discussion.

Chapitre 2

Etat de l'art : la voix et ses modifications dans MP

2.1 Le son : origine, capture et transmission téléphonique

2.1.1 Origine et propagation du son

Le son est une vibration mécanique d'un fluide qui se propage sous la forme d'ondes progressives longitudinales grâce à la déformation élastique de ce fluide.

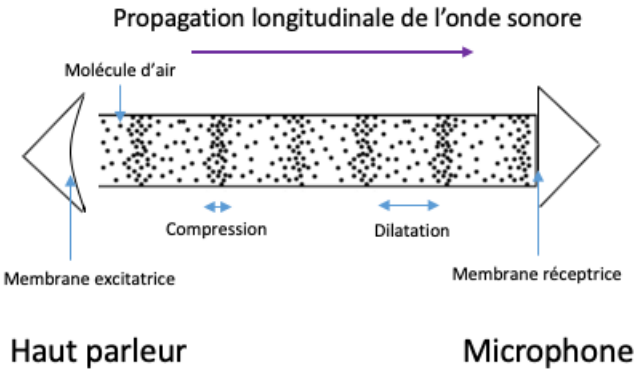


FIGURE 2.1 – Propagation du son dans l'air. La membrane du haut-parleur en vibrant entraîne une succession de compression et dilatation de la couche d'air à proximité. Cette dernière se met alors à osciller, propageant ainsi l'onde de pression à la couche suivante et ainsi de suite jusqu'à la membrane du microphone.

Sous l'effet d'une excitation mécanique, par exemple la vibration d'une membrane (de haut-parleur, de tambour), les molécules du fluide à proximité vont osciller autour de leurs positions initiales créant une succession de compressions et dilatations qui vont entraîner une oscillation de la couche suivante et ainsi de suite. L'onde de pression ainsi créée va se propager et peut être captée par l'oreille ou un microphone en entraînant la vibration de la membrane réceptrice (tympan ou membrane du microphone), qui va elle-même engendrer un signal électrique. Définissons le signal sonore p(t) comme la pression acoustique (variation de la pression atmosphérique) au niveau du capteur, ce qui revient à considérer l'écart à l'équilibre de la membrane de ce dernier.

2.1.2 Capture du son

Le traitement du signal sonore nécessite l'enregistrement de l'onde sonore et sa numérisation. Les microphones convertissent l'onde de pression acoustique en signal électrique analogique, qui est lui-même converti dans un deuxième temps en signal numérique.

Microphones Les microphones sont des transducteurs électroacoustiques, c'est-à-dire des appareils qui convertissent un signal acoustique en un signal électrique. Ils peuvent capter directement la pression, son gradient ou les deux. Il existe plusieurs manières pour les microphones de convertir la vibration de leur membrane en signal électrique.

Les premiers microphones étaient des microphones à charbon. Ils utilisaient une poudre granuleuse de carbone qui avait une résistance variable quand elle était comprimée.

Les microphones dynamiques fonctionnent grâce au mécanisme d'induction électromagnétique. Pour les microphones dynamiques à bobine mobile, le mouvement de la membrane entraîne le déplacement d'une bobine placée dans un aimant. Ce déplacement d'un conducteur dans un champ magnétique entraîne par induction un courant électrique. Les microphones à ruban fonctionnent sur le même principe sauf qu'un fin ruban métallique remplace la bobine, faisant aussi lieu de membrane. L'avantage des microphones dynamiques est leur robustesse (pour les microphones à bobine), le fait qu'ils ne nécessitent pas d'alimentation externe et leur prix plus abordable que les microphones électrostatiques.

Enfin les microphones électrostatiques fonctionnent grâce au principe du condensateur. Un condensateur, dont l'une des armatures est la membrane du microphone est connecté à un générateur et à une résistance. En se déplaçant la membrane entraîne une variation de capacité du condensateur qui produit une variation de tension dans le courant électrique. Les microphones électrostatiques ont besoin d'une alimentation externe pour la polarisation du condensateur et pour une préamplification du signal. Le microphone à électret est une sous-catégorie de microphones électrostatiques. Un matériau avec une charge électrostatique permanente fait guise d'armature, l'autre armature du condensateur étant toujours la membrane du micro. Il ne nécessite plus de polarisation externe mais requiert quand même une alimentation externe pour la préamplification. Les microphones électrostatiques sont les plus utilisés par les professionnels du son, car ils ont généralement un meilleur rapport signal sur bruit et une réponse en fréquence plus large et plus plate, donc déformant peu le signal.

Les microphones implantés dans les téléphones ont longtemps été des microphones à charbon. Maintenant ils sont remplacés par des variantes du microphone à électret, c'est le cas des MEMS (Microsystèmes électromécaniques) pour les smartphones.

L'autre caractéristique importante d'un microphone est sa directivité, c'est à dire sa sensibilité à la direction d'où provient le son.

Le microphone omnidirectionnel capte le son de façon uniforme selon toutes les directions. Ce sont des capteurs de pression. Ils sont sensibles à la réverbération mais peu aux bruits de manipulation et au vent et ne déforment pas le timbre.

Le microphone bi-directionnel (ou en 8) est un capteur de gradient de pression qui reçoit le son devant et derrière mais pas sur les côtés. La plupart des microphones à ruban sont bi-directionnels.

Les microphones unidirectionnels privilégient les sources placées devant le micro. Ils résultent d'une association du type capteur de pression et capteur de gradient de pression. Il existe les directivités cardioïdes (la plus répandue), sous cardioïdes, super-cardioïdes, hypercardioïdes et canon (forte directivité vers l'avant) cf. Figure 2.2 . Ils sont peu sensibles aux bruits extérieurs lointains mais plus sensibles aux bruits de manipulation, au vent et aux plosives. La voix est également légèrement déformée avec un effet de proximité (les graves sont amplifiés quand le micro est proche de la source), et un effet de distorsion.

FIGURE 2.2 – Directivité des microphones. De gauche à droite : omnidirectionnel, bi-directionnel, sous-cardioïd, cardioïd, hypercardioïd, supercardioïd et canon. Source : https ://en.wikipedia.org/wiki/Microphone

Numérisation Les microphones produisent un signal électrique analogique (continu) à partir des vibrations mécaniques de leur membrane. Or le traitement informatique se fait sur des signaux numériques (valeurs discrètes binaires). Ainsi une conversion analogique numérique est effectuée sur le signal électrique produit par le microphone. La méthode la plus couramment utilisée pour convertir un signal vocal analogique en numérique est la méthode PCM (pour *Pulse Code Modulation*).

Elle est composée de trois étapes : l'échantillonnage, la quantification et l'encodage. Le traitement de chaque échantillon se fait sans chiffrement et sans compression de données. La qualité du signal numérique dépendra de la fréquence d'échantillonnage et du nombre de bits (valeurs binaires) qui sont attribués au codage de chaque valeur extraite.

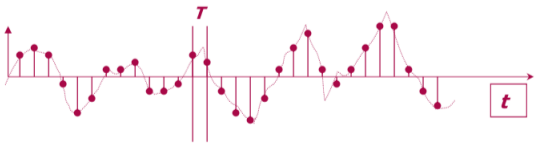


FIGURE 2.3 – Echantillonnage d'un signal x(t) à la fréquence d'échantillonnage $1/T$. Le signal à valeurs continues dans le temps est transformé en signal à valeurs discrètes $x(n) = x(nT)$. Source : [Richard, 2016].

Échantillonnage La fréquence d'échantillonnage est la fréquence à laquelle les valeurs sont extraites. Si cette fréquence est très faible cela entraînera une grande perte d'information, si la fréquence est très élevée, il y aura moins de perte d'information mais cela nécessitera plus d'espace de stockage. La fréquence d'échantillonnage doit valoir au minimum le double de la fréquence maximale présente dans le signal d'après le théorème de Nyquist-Shannon. Les fréquences d'échantillonnage couramment utilisées vont de 96kHz pour les analyses audio nécessitant beaucoup de précision à 8kHz pour le téléphone (minimum nécessaire pour couvrir la bande passante de 160Hz à 3500Hz permettant une transmission de la parole suffisamment intelligible).

Quantification et encodage La phase de quantification consiste à arrondir les valeurs réelles prises aux points d'échantillonnage à une valeur prise dans un ensemble fini. Les valeurs possibles peuvent résulter d'une quantification linéaire, ou non linéaire (logarithmique pour le téléphone par exemple). Ensuite un encodage convertit cette valeur en un code unique, souvent un nombre entier exprimé en binaire. Habituellement 8 ou 16 bits sont utilisés par valeur, ce qui permet de coder respectivement 2^8 ou 2^{16} valeurs différentes. Dans le cas de 8 bits, les codes vont de 0 à 255, ou de -128 à 127 (on parle dans ce cas de PCM signé). La transposition en binaire des valeurs signées se fait soit avec un bit de signe au début, soit en complément à 255, c'est à dire que les valeurs positives sont codées par les transpositions binaires de 0 à 127 et les valeurs négatives par transposition binaire des valeurs 128 à 255. Quand plus de 2 octets (soit

16 bits) sont utilisés par valeur, il faut préciser l'ordre d'importance des octets, on parle alors de *big endian* (BE) quand les octets les plus significatifs sont au début et de *little endian* (LE) quand les octets les plus significatifs sont à la fin. Ainsi un PCM S24 LE est un PCM signé avec une précision de 24 bits et little endian.

2.1.3 Transmission du son : téléphonie

2.1.3.1 Les réseaux de téléphonie

Jusque dans les années 70, la transmission de la voix d'un téléphone à un autre, ou d'un téléphone à un magnétophone, se faisait intégralement en mode analogique.

Dans les années 80, les opérateurs téléphoniques ont commencé à numériser les signaux voix entre les centraux téléphoniques, dans deux buts : améliorer les performances du multiplexage (transmissions de signaux différents en même temps via un même support), et faciliter la transmission des signaux au sein des centraux intermédiaires. Actuellement la plupart des téléphones filaires fonctionnent sur ce principe.

Dans les années 90, la conversion en signal numérique a commencé à se faire à l'intérieur même de certains types de téléphones. C'est le cas des téléphones portables qui utilisent le réseau GSM (*Global System for Mobile Communications*) pour transmettre de façon numérique la voix du téléphone à l'antenne relai.

De nos jours, si les opérateurs téléphoniques utilisent encore des canaux dédiés pour assurer la qualité de la transmission téléphonique au sein de leurs réseaux, l'évolution des ratios performances/prix dans la transmission de données permet de transporter la voix sur des couches de réseau de type Internet. C'est ce qu'on appelle le VoIP (Voix sur IP - *Voice over Internet Protocol*).

Le réseau IP est au départ prévu pour transporter des données. Au moment de la transmission, les données sont découpées en paquets numérotés, ces paquets sont transmis au sein de réseau par le meilleur chemin possible – qui peut évoluer en fonction de la charge à tout instant, ce qui conduit les paquets à pouvoir prendre des chemins différents – et à la réception, les paquets sont remis dans l'ordre. Quand on transmet de la voix par IP, on utilise un protocole spécial, le RTP (*Real Time Transport Protocol*), optimisé pour diminuer la latence entre l'émission et la réception des paquets, et ainsi permettre une communication en temps réel. Ce protocole est toujours associé à un autre protocole, comme le protocole SIP (*Session Initiation Protocol*) qui gère l'établissement et la fin de la session.

Certains téléphones fixes, en entreprise par exemple, utilisent la VoiP à la place du réseau RTC (Réseau Téléphonique Commuté). De nombreux logiciels (Skype, Messenger, What's app...) disponibles sur ordinateur ou smartphone utilisent la VoIP pour transmettre la voix et ou de la vidéo.

2.1.3.2 Encodage

De manière à diminuer les coûts liés au transport du signal, les signaux une fois numérisés connaissent ensuite une étape d'encodage complémentaire avec compression. Les codecs (programmes responsables de l'encodage et du décodage) les plus utilisés en téléphonie sont les suivants :

Codec à bande étroite : Les codecs à bande étroite encodent les signaux en 8kHz avec une bande de fréquence comprise entre environ 300 et 3400 Hz.

- G.711 est le principal codec utilisé sur le réseau RTC et peut être utilisé en VoIP. Cet encodage consiste à transformer une quantification PCM linéaire (sur 13 ou 14 bits) en quantification de type logarithmique sur 8 bits. Il y a deux versions légèrement différentes du G.711

le G.711μ pour l'Amérique du Nord et le Japon, et le G.711A pour la plupart des autres pays (dont la France).

- AMR (Adaptive Multi Rate) est le système d'encodage utilisé sur le réseau de téléphonie mobile GSM et ses dérivés (UMTS). Le signal sonore est d'abord encodé en 8kHz sur 13 bits linéaires avec une bande passante de (300-3400 Hz). Il est ensuite compressé selon 9 modes possibles : 8 modes de transmission de parole (allant de 4.75 à 12.20 kbit/s), et un mode de transmission de bruit de fond (à 1.80 kbit/s). La compression s'effectue de manière dynamique, pouvant passer d'un mode à l'autre toutes les 20ms, suivant le contenu audio et la qualité de la transmission.

Codec à large bande : Les codecs à large bande sont uniquement utilisés en VoIP. Ils permettent la transmission des signaux avec une bande de fréquence plus large (50-7000 Hz) avec une fréquence d'échantillonnage de 16KHz et une résolution de 8 à 16 bits. Les codecs associés les plus courants sont le G.722, le G.722.2 et le G.729.1.

2.1.3.3 Limitations

La transmission du son par téléphonie connaît un certain nombre de limitations qui détériorent plus ou moins le signal :

- **Fréquence d'échantillonnage et bande de fréquences réduites**, surtout pour la téléphonie en bande étroite.
- **Paquets manquants** (ou perdus) : quand les mémoires tampons sont saturées, certains paquets peuvent ne plus être stockés et donc ne seront pas transmis. De même un paquet arrivant avec trop de retard sera considéré comme manquant. Les paquets manquants sont caractéristiques des transmissions numériques, ils ne concernent pas les transmissions analogiques. Le codec G.711 garantit normalement une perte maximale de 1% des paquets.
- **Latence** : c'est le temps de transmission de la voix. La latence est plus importante dans les réseaux IP que RTC. Elle est considérée comme acceptable (d'après les recommandations UIT-T G114) si elle est inférieure à 200ms.
- **Gigue** : c'est la variation de latence par rapport à la moyenne. Elle peut être définie comme étant l'écart type de la latence, ou la différence entre la latence maximale et la moyenne des latences. Un *buffer* de gigue à la réception permet de compenser la gigue en retenant les paquets pendant un certain temps avant de les jouer. Si un paquet arrive avec trop de retard, le *buffer* ne suffit pas, et le paquet est considéré comme perdu.
- **Distorsion d'amplitude et de phase** : déformation du signal pendant les parties de transmission analogiques du réseau.
- **Bruit de fond** : présence d'un bruit de fond additif lors de la transmission analogique (dû autre autre au mouvement brownien des électrons). Présence d'un bruit de fond (plus faible) de numérisation pour les transmissions numériques (dû à l'étape de quantification).
- **Distorsion d'amplitude due aux codecs** : source de déformations supplémentaires.
- **Echo** : la gêne due à l'écho dépend de l'amplitude de celui-ci et de la latence. Les appels RTC nationaux (latence < 25 ms) ne provoquent généralement pas d'écho, contrairement aux appels RTC internationaux ou sur réseau IP qui requièrent des annulateurs d'écho.

2.2 Analyse du son et de la voix

2.2.1 Analyse du son

2.2.1.1 Caractéristiques générales du son

L'oreille humaine entend les sons produits entre 20Hz et 20kHz. Les sons supérieurs à 20kHz sont dits ultrasons et les sons inférieurs à 20Hz infrasons. Un son peut être représenté dans le domaine temporel (amplitude du signal dans le temps) ou dans le domaine fréquentiel (visualisation de sa composition fréquentielle). La représentation du son dans le domaine fréquentiel s'appelle le spectre du signal. Le son peut être décrit suivant trois caractéristiques principales : sa hauteur, son timbre et son intensité.

Hauteur Si le son est harmonique, c'est-à-dire qu'il contient des fréquences multiples d'une fréquence fondamentale audible, cette dernière détermine la hauteur tonale d'un son. Plus la fréquence fondamentale est élevée, plus le son est aigu. L'oreille est sensible au log de la fréquence, ainsi un saut d'une octave correspond à une multiplication de la fréquence par deux.

Timbre Le timbre est la caractéristique qui permet le mieux d'identifier la source d'un son. Il est décrit par la répartition des fréquences dans le spectre sonore, autrement dit par l'enveloppe spectrale du son. Quand le son est harmonique le timbre caractérise les différents poids des harmoniques.

Intensité L'intensité acoustique J d'une onde sonore progressive est définie comme la puissance qu'elle transporte par unité de surface (en $W.m^{-2}$).

$$J = \frac{puissance\ acoustique}{surface} \tag{2.1}$$

Ce qui équivaut à :

$$J = \frac{Peff^2}{Z} \tag{2.2}$$

avec $Peff = \sqrt{< p(t)^2 >}$ la pression acoustique efficace et Z l'impédance acoustique du milieu.

L'oreille humaine est également sensible au log de l'intensité acoustique. C'est pourquoi en pratique on utilise plutôt le niveau d'intensité acoustique (SLP pour *Sound Level Pressure*) défini par :

$$I = 10 * log_{10} \frac{J}{J_0} \tag{2.3}$$

avec $J_0 = 10^{-12} W m^{-2}$ le seuil d'audibilité dans l'air de l'oreille humaine pour un son sinusoïdal de 1 kHz. I s'exprime en décibel (dB). Une intensité deux fois plus forte correspond à une augmentation de 3 dB, ce qui correspond aussi à la sensibilité minimale de l'oreille. Une conversation entre deux personnes est de l'ordre de 50 dB et le seuil de douleur est atteint autour de 120 dB.

2.2.1.2 Aspect ondulatoire

Suivant la composition fréquentielle du son, on peut distinguer les sons purs (sons sinusoïdaux), les sons périodiques (ou harmoniques), les sons quasi-périodique et les sons apériodiques (bruits).

Son pur Le signal s(t) d'un son pur est représenté par une onde sinusoïdale dans le domaine temporel :

$$s(t) = Asin(\omega t + \varphi) \tag{2.4}$$

avec A l'amplitude, ω la pulsation et φ la phase.

L'intensité acoustique vaut alors :

$$I = \frac{A^2}{2Z} \tag{2.5}$$

et la fréquence fondamentale (et unique fréquence ici) :

$$Fo = \frac{\omega}{2\pi} \tag{2.6}$$

Dans le domaine fréquentiel, ce signal apparaît comme une raie à l'abscisse correspondant à la fréquence fondamentale cf. Figure 2.4.

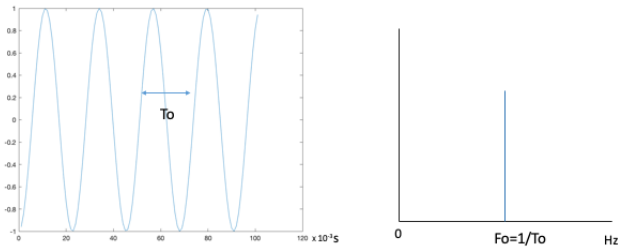


FIGURE 2.4 – Signal d'un son pur sinusoïdal de fréquence Fo, représentation temporelle (à gauche) et fréquentielle (à droite)

L'exemple le plus connu d'émetteur de son pur (ou quasiment) est le diapason.

Son périodique Les sons purs exposés précédemment se rencontrent peu dans la nature. La plupart des sons sont en réalité composés d'une somme (continue ou discrète) de sons purs.

Les signaux sonores périodiques de période T peuvent être décomposés en série de Fourier, c'est à dire en une somme de signaux sinusoïdaux dont les fréquences sont des multiples de la fréquence fondamentale $1/T$.

$$s(t) = \sum_{n=1}^{+\infty} a_n cos(n\omega t + \varphi_n) \tag{2.7}$$

$a_n cos(n\omega t + \varphi_n)$ est défini comme l'harmonique de rang n. L'harmonique de rang 1 est appelée la fondamentale, $\omega/2\pi$ étant la fréquence fondamentale ($= 1/T$). Les amplitudes a_n des harmoniques tendent vers 0 lorsque n tend vers l'infini.

Dans le domaine fréquentiel les sons périodiques (ou harmoniques) sont représentés en un ensemble de raies avec pour abscisse les fréquences des harmoniques et pour ordonnée l'amplitude (le poids) a_n de celles-ci, ou l'intensité correspondante (en dB). Les harmoniques étant des multiples de la fréquence fondamentale, l'écart entre deux harmoniques consécutives est égal à cette dernière.

La hauteur du son est définie par la fréquence fondamentale. Cette dernière correspond à la périodicité des raies. On notera que même si $a_1 = 0$ (c'est à dire un poids nul pour la fréquence fondamentale) la hauteur du son perçue sera celle de la fréquence fondamentale.

Les sons harmoniques peuvent être perçus dans le domaine temporel comme le produit de convolution d'une source et d'un filtre (ou résonateur), ce qui correspond à leur produit dans le

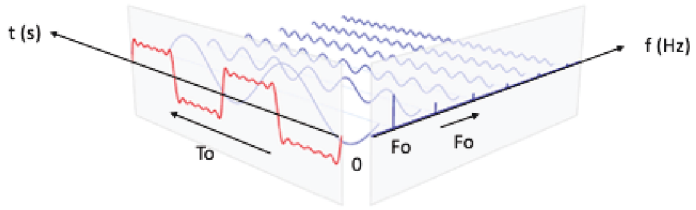


FIGURE 2.5 – Décomposition d'un signal périodique en signaux sinusoïdaux (harmoniques). Représentation temporelle à gauche et spectrale à droite. L'écart entre les raies du spectre correspond à la fréquence fondamentale $Fo = 1/To$ avec To la période du signal.

domaine fréquentiel (cf. Figure 2.6). La source correspond à la production aux vibrations (cordes d'une guitare, cordes vocales..). Elle génère les ondes sinusoïdales correspondant à la fréquence fondamentale et aux harmoniques. Elle va déterminer la hauteur du son. Ces ondes vont ensuite passer dans un résonateur (caisse de résonance d'une guitare, conduit vocal ...) qui va jouer le rôle de filtre. Il va amplifier certaines harmoniques et en réduire d'autres. Le résonateur va donner l'enveloppe spectrale, qui est responsable du timbre du son.

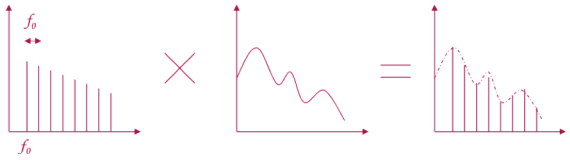


FIGURE 2.6 – Le spectre d'un son harmonique peut être considéré comme le produit du spectre d'une source et du spectre d'un résonateur (filtre). La source produisant la fréquence fondamentale et ses harmoniques, elle est responsable de la hauteur du son. Le résonateur génère l'enveloppe spectrale en amplifiant et réduisant certaines harmoniques, il est responsable du timbre. Adapté de [Richard, 2016].

Son quasi-périodiques et apériodique Les sons harmoniques purs sont également peut fréquents, ils sont souvent accompagnés d'un bruit plus ou moins prononcé, ils sont dits quasi-périodiques (cf. Figure 2.7a). Certains sons ne présentent pas d'aspect cyclique dans leur décours temporel, ils sont dits apériodiques (cf. Figure 2.8a). Les sons quasi-périodiques et apériodiques peuvent être considérés comme une somme continue (une intégrale) de sons élémentaires sinusoïdaux. Les amplitudes des composantes sinusoïdales sont obtenus par Transformation de Fourier (TF) du signal temporel.

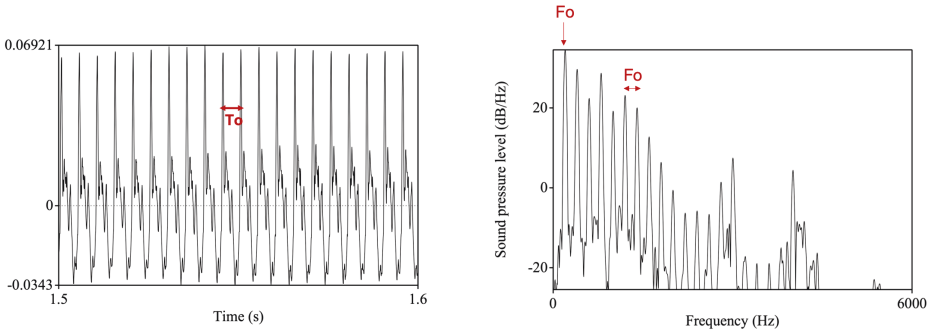
$$s(t) = 1/2\pi \int_{-\infty}^{+\infty} \hat{s}(x) e^{ixt} dx = TF^{-1}(\hat{s}) \tag{2.8}$$

avec

$$\hat{s}(x) = TF(s) = \int_{-\infty}^{+\infty} s(\xi) e^{ix\xi} d\xi \tag{2.9}$$

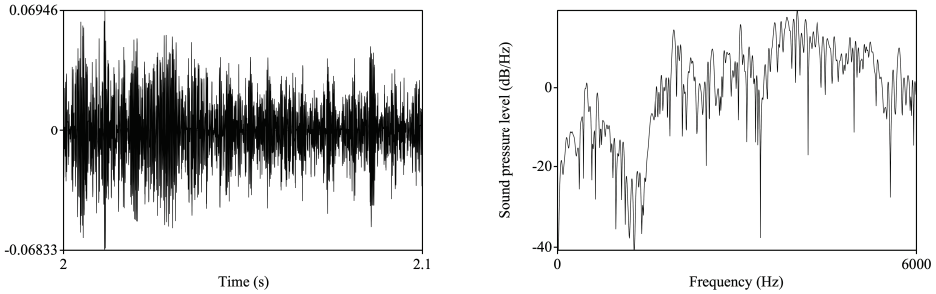
La représentation fréquentielle illustre la densité spectrale de puissance (aussi appelée périodogramme), soit le carré du module de la transformée de Fourier. On peut aussi représenter la densité spectrale d'amplitude avec directement le module de la TF. Pour les sons quasi-périodiques, la structure est plus continue que pour les sons périodiques, avec des pics au niveau

des harmoniques (cf. Figure 2.7a). Pour les sons apériodiques la représentation spectrale est d'allure continue sans pics périodiques.



(a) Représentation temporelle (b) Représentation fréquentielle

FIGURE 2.7 – Représentation temporelle et fréquentielle d'un son quasi-périodique (le phonème $/a/$). Dans la représentation temporelle, on constate que le signal est quasi-périodique avec une période $To = 5ms$. Dans la représentation spectrale on remarque la présence de pics (correspondant aux harmoniques) espacés de $Fo = 200Hz$ (la fréquence fondamentale). Fo est aussi la fréquence du premier pic et sa valeur vaut bien $1/To$.



(a) Représentation temporelle (b) Représentation fréquentielle

FIGURE 2.8 – Représentation temporelle et fréquentielle d'un son apériodique (le phonème $/ch/$). On ne note pas de période dans le signal temporel, ni la présence de pics régulièrement espacés dans la représentation spectrale.

En traitement du signal sonore, l'outil le plus utilisé pour générer les spectres est la transformation de Fourier rapide. C'est un algorithme de calcul de la transformation de Fourier discrète (équivalent discret de la transformée de Fourier, que l'on peut appliquer aux signaux numériques).

2.2.1.3 Aspect dynamique : le spectrogramme

Les caractéristiques d'un son (fréquence fondamentale, intensité, timbre) peuvent varier avec le temps. Cette évolution avec le temps est décrite dans la représentation temporelle mais pas dans la représentation fréquentielle comme décrite ci-dessus. Lorsque le son n'est pas stationnaire, il est alors découpé en segments (qui peuvent se chevaucher) pendant lesquelles il est considéré comme stationnaire. Pour effectuer une représentation fréquentielle de la voix, on découpe le signal sonore en segments de quelques dizaines de ms. On considère que pendant cette durée la voix n'a pas le temps de trop se modifier et donc que le son est stationnaire. Un spectre peut alors être calculé pour chaque segment. L'évolution de ces spectres en fonction

du temps peut être visualisée par un spectrogramme (cf. Figure 2.9). C'est une représentation temps-fréquence où pour chaque fenêtre temporelle le spectre calculé est représenté verticalement. Une échelle de couleur, ou de nuances de gris représente les poids des fréquences du spectre.

La taille des segments (ou fenêtres) peut être ajustée. Des fenêtres temporelles de courte durée (inférieure à la période T_0) donneront une bonne précision temporelle mais une moins bonne précision fréquentielle, cette résolution est appelée à bandes larges (cf. Figure 2.10b). A l'inverse, des fenêtres temporelles de plus grande durée (supérieure à $2T_0$) impliqueront une moins bonne résolution temporelle mais une meilleure résolution spectrale, ce réglage est dit à bandes étroites (cf. Figure 2.10a). Le réglage à bandes étroites permet une bonne visualisation de la fréquence fondamentale et de ses harmoniques, tandis que le réglage à large bande est plus adapté pour visualiser le timbre du son.

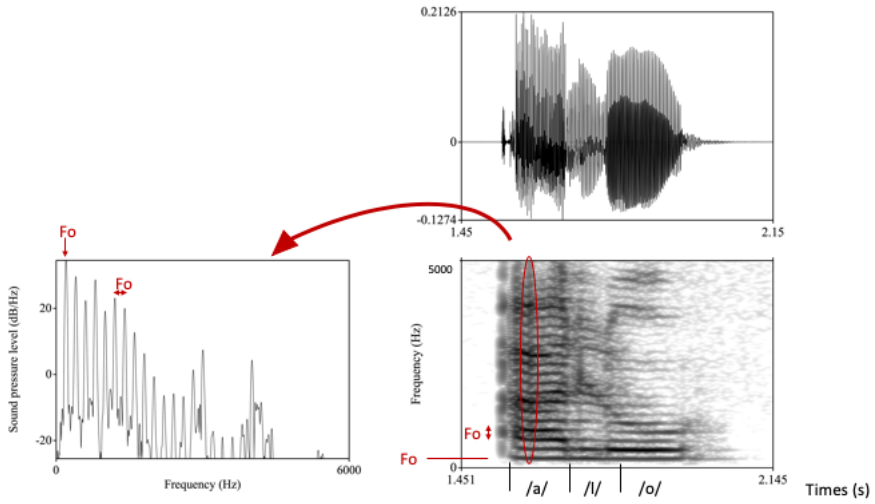
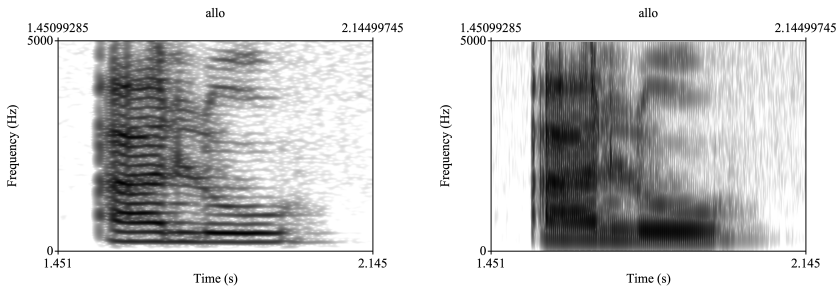


FIGURE 2.9 – Représentation du spectrogramme (en bas à droite) et du signal temporel (en haut) correspondant au mot /allo/. Pendant le /a/ et le /o/ on observe des raies correspondant à la fréquence fondamentale et ses harmoniques. A gauche figure le spectre correspondant au segment entouré.



(a) Réglage bandes étroites (b) Réglage larges bandes

FIGURE 2.10 – Spectrogrammes correspondant au mot /allo/ avec un réglage en bandes étroites (à gauche) et en bandes larges (à droite). On constate que le réglage à bandes étroites est plus adapté pour visualiser la fréquence fondamentale et ses harmoniques, tandis que le réglage à large bande est plus adapté pour visualiser le timbre.

2.2.2 Phonétique

Nous allons maintenant nous concentrer sur les sons produits par la voix humaine, l'étude de ces sons est appelée phonétique. La parole est constituée de son quasi-stationnaire, de son apériodique (bruits) et de silences. Le spectre de fréquences de la voix s'étend de 50Hz à 20kHz environ.
Cette partie utilise le cours de [Gabriel, 2019].

2.2.2.1 Production de la parole

La production de la parole nécessite la coordination d'une centaine de muscles. Les organes impliqués dans la production de la voix peuvent être divisés en trois parties : l'appareil respiratoire (les poumons), le larynx (les cordes vocales) et le conduit vocal (cavités pharyngale, orale et nasale).

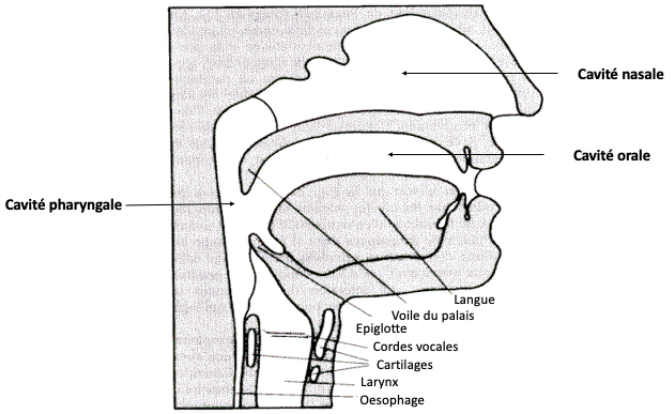


FIGURE 2.11 – Organes impliqués dans la production de la voix.

Appareil respiratoire ou partie sub-glottique La majorité des sons des langues du monde utilisent le système respiratoire. L'inspiration remplit les poumons d'air grâce à la contraction du diaphragme. La rétraction de ce muscle comprime les poumons qui expulsent l'air par les bronches et la trachée vers la gorge. Durant l'activité de parole, la respiration est modifiée. Sa durée et son intensité sont contrôlées de manière à envoyer la quantité d'air nécessaire au fonctionnement de la production sonore.

Le larynx ou partie glottique Une fois expulsé par les poumons, l'air entre dans le larynx. Pendant la respiration libre les cordes vocales (plis vocaux situés dans le larynx) sont relâchées, les membranes sont alors écartées. L'écart entre les cordes vocales est appelé fente glottique ou glotte. Pendant la production de certains sons, les cartilages du larynx ferment les cordes vocales. Lorsque le flux d'air issu des poumons arrive à leur niveau, la pression sous les cordes entraîne leur ouverture, une partie du flux d'air s'écoule, ce qui crée une diminution de la pression sub-glottique et donc la fermeture des cordes vocales et ainsi de suite. La fréquence d'ouverture et de fermeture des cordes vocales dépend de leur longueur, leur épaisseur et de leur tension. Elle est d'environ 100Hz chez l'homme, 180 Hz chez la femme et 400Hz chez l'enfant. La vibration des cordes vocales crée une onde sonore quasi périodique riche en harmoniques dont la fréquence fondamentale correspond à la fréquence de vibration des cordes vocales et donne la hauteur de la voix (appelée pitch). Le son produit est dit voisé. L'intensité du son dépend de l'amplitude des vibrations des cordes vocales. Cette amplitude dépend de la tension des cordes, de la quantité d'air cherchant à passer (donc de la respiration) et du degré de fermeture des cordes. Si la fermeture est incomplète, une partie du flux d'air est expirée sans être transformée en vibration, produisant un son de plus faible intensité et bruité.

Conduit vocal ou partie supra-glottique Le flux d'air sortant des cordes vocales en vibration passe ensuite par le conduit vocal, où certaines harmoniques vont être amplifiées et d'autres atténuées, selon la forme et le volume de ce dernier. Le conduit vocal est composé de trois cavités (pharyngale, orale et nasale) de formes et tailles variables qui vont donner son timbre au son voisé, via le phénomène de résonance. Pour rappel, un système résonant (comme une cavité), lorsqu'il est soumis à une onde excitatrice, va amplifier les composantes fréquentielles proche de sa fréquence caractéristique (aussi appelée fréquence de résonance ou propre) et amortir les autres composantes fréquentielles. La fréquence caractéristique d'une cavité dépend de sa forme et de son volume. Plus la cavité est grande plus sa fréquence propre est basse. Lorsque plusieurs cavités sont reliées entre elles, un phénomène de couplage modifie leurs fréquences de résonance. Les articulateurs du conduit vocal peuvent modifier la forme, le volume et le couplage des cavités vocales, permettant une modulation du timbre du son et déterminant la forme du spectre émis.

La cavité pharyngale est la première cavité par laquelle le son sortant des cordes vocales passe. C'est un carrefour aéro-digestif situé entre le larynx et les fosses nasales d'une part et entre l'œsophage et la bouche d'autre part. Sa paroi est constituée de muscles constricteurs qui permettent d'en modifier son diamètre. Son volume peut également être modifié par le recul ou l'avancement de la racine de la langue, ainsi que par l'abaissement du larynx.

L'air passe ensuite dans la cavité orale. On peut la séparer en deux cavités : la cavité buccale (située entre les joues, les dents et le palais) et la cavité labiale qui se forme lorsque les lèvres sont projetées en avant, produisant alors des sons dits labialisés. Les mouvements de la langue, des lèvres et de la mâchoire y induisent de nombreux changements de configuration, en augmentant par exemple le volume du conduit vocal ou en l'obstruant. La cavité orale est la partie du tractus vocal la plus importante dans l'articulation car c'est celle qui connaît le plus de modifications lors de la production de la parole.

Lorsque le voile du palais est baissé, l'air issu du larynx passe également dans la cavité nasale. Cette dernière est constituée de deux cavités séparées par une cloison verticale. Cette cavité a des dimensions et une forme fixes. Les sons alors produits sont dits nasaux.

Lorsque l'air passe par les cordes vocales ouvertes sans les faire vibrer, un son dit non voisé est possible par obstruction du conduit vocal entraînant un écoulement turbulent du flux d'air.

2.2.2.2 Sons vocaliques et sons consonantiques

En linguistique le phonème est le plus petit élément qu'on peut isoler dans la chaine parlée. Un phonème est en réalité une entité abstraite pouvant correspondre à plusieurs sons, dépendant du locuteur ou de sa position dans le mot. On appelle phones les différentes réalisations d'un phonème. On compte 36 phonèmes dans la langue française, que l'on peut séparer en 16 voyelles, 17 consonnes et 3 semi-voyelles (ou semi-consonnes) (cf. Tableau 2.1). Quelques consonnes d'emprunts aux langues étrangères sont parfois rajoutées comme : le phonème /h/ (de hop) et phonème /ŋ/ (de camping).

La distinction entre voyelles et consonnes se fait classiquement sur des critères articulatoires : la production des voyelles suppose une libre circulation de l'air à partir de la glotte alors que la production des consonnes nécessite une obstruction partielle ou complète en un ou plusieurs endroits du conduit vocal. La différence entre voyelles et consonnes peut aussi être définies d'un point de vue linguistique, de par leur place dans la syllabe. Les voyelles peuvent être le noyau de la syllabe au contraire des consonnes. Les semi-voyelles (ou semi-consonnes) présentent le même aspect articulatoire que les voyelles mais ont la même place que les consonnes dans la syllabe.

Voyelles Les voyelles sont des sons caractérisés par le libre passage de l'air dans les cavités supra-glottiques. Ce sont des sons voisés qui suivent le modèle source-filtre. La fréquence fondamentale et les harmoniques sont données par les cordes vocales (source). La fréquence fonda-

voyelles	consonnes	semi-consonnes
/i/ : il	/p/ : père	/j/ : yeux
/e/ : blé	/t/ : terre	/w/ : oui
/ɛ/ : colère	/k/ : cou	/ɥ/ : lui
/a/ : patte	/b/ : bon	
/ɑ/ : pâte	/d/ : dans	
/ɔ/ : mort	/g/ : gare	
/o/ : chaud	/f/ : feu	
/u/ : genou	/s/ : sale	
/y/ : rue	/ʃ/ : chat	
/ø/ : peu	/v/ : vous	
/œ/ : peur	/z/ : zéro	
/ə/ : le	/ʒ/ : je	
/ɛ̃/ : plein	/l/ : lent	
/ɑ̃/ : sans	/ʁ/ : rue	
/ɔ̃/ : bon	/m/ : main	
/œ̃/ : brun	/n/ : nous	
	/ɲ/ : agneau	

TABLE 2.1 – Liste des phonèmes de la langue française

mentale est généralement comprise entre 80Hz et 400Hz et les dernières harmoniques détectées peuvent aller jusqu'à 10kHz. Le timbre est quant à lui formé par les cavités résonantes du conduit vocal (filtre). Le timbre dépend essentiellement du nombre de cavités traversées par le son (la cavité labiale et la cavité nasales ne sont pas toujours impliquées), de la forme et du volume de la cavité buccale.

Le nombre de cavités traversées permet une distinction entre :
— les voyelles nasales (présence du résonateur nasal) et les voyelles orales (absence de ce dernier) ;
— les voyelles arrondies (présence du résonateur labial) et les voyelles non arrondies (absence de ce dernier).

La forme de la cavité buccale, qui dépend essentiellement du point d'articulation (positionnement de la langue), va entraîner une distinction entre les voyelles antérieures (la partie avant de la langue se rapproche de l'avant du palais), les voyelles postérieures (l'arrière de la langue se rapproche de l'arrière du palais) et les voyelles centrales (la partie centrale de la langue se rapproche du palais).

Le volume de la cavité buccale, qui dépend principalement du degré d'ouverture de la bouche, permet la distinction entre les voyelles fermées, mi fermées, mi ouvertes et ouvertes (dans l'ordre croissant d'ouverture).

Le classement des voyelles selon ces critères est souvent représenté par un triangle (ou trapèze) vocalique, avec pour axe horizontal la profondeur du point d'articulation et pour axe vertical le degré d'ouverture de la bouche, cf. Figure 2.13. Les voyelles se situant aux extrémités de ce triangle sont les phonèmes /*a*/ (produit avec une grande ouverture), /*i*/ (produit avec une faible ouverture et une articulation antérieure) et /*u*/ (produit avec une faible ouverture et une articulation postérieure).

	Antérieure		Centrale	Postérieure	
	Non arrondie	Arrondie		Non arrondie	Arrondie
Fermée	i	y			u
Mi-fermée	e	ø			o
Moyenne			ə		
Mi-ouverte	ɛ • ɛ̃	œ • œ̃			ɔ • ɔ̃
Ouverte	a			ɑ • ɑ̃	

FIGURE 2.12 – Les voyelles du français. Lorsque deux voyelles apparaissent par paire, celle de gauche correspond à la voyelle orale et celle de droite à la voyelle nasale

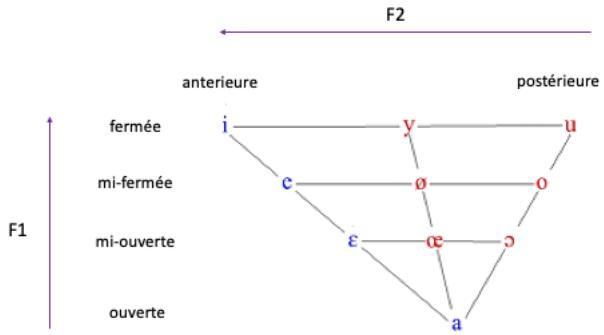


FIGURE 2.13 – Triangle vocalique représentant un classement des voyelles orales françaises. L'axe horizontal représente le point d'articulation et l'axe vertical le degré d'ouverture. Le premier formant F1 et deuxième formant F2 corrèlent respectivement avec les degrés d'ouverture et le lieu d'articulation.

Les fréquences de résonance du conduit vocal lors de la production de voyelles sont appelées formants. Ils peuvent être visualisés sur le spectre : ce sont les pics de l'enveloppe spectrale. On peut également les visualiser sur le spectrogramme avec un réglage en bandes étroites : ils sont situés au niveau des bandes foncées (cf. Figure 2.15).

Le premier formant est déterminé par le degré d'ouverture de la bouche et la hauteur de la langue. Le deuxième formant est la résultante de la position de la langue (avant ou arrière) et de la forme des lèvres (étirées ou arrondies). Le troisième formant dépend juste de la forme des lèvres. Les valeurs des trois premiers formants des voyelles orales sont détaillées dans la Figure 2.14. Les voyelles nasales présentent des caractéristiques plus complexes. Le couplage supplémentaire avec la cavité nasale crée un formant supplémentaire, appelé formant nasal (autour de 600Hz) ainsi que des "anti-formants" (zones où on observe une forte diminution de l'intensité des harmoniques). Le couplage modifie également les valeurs des formants correspondant aux voyelles équivalentes orales.

Il est important de noter que des configurations différentes du tractus vocal peuvent donner un même résultat acoustique et donc un même phonème. Comme expliqué dans [Ghio and Pinto, 2007], un arrondissement des lèvres peut par exemple être compensé par une avancée du point d'articulation. L'interprétation articulatoire des résultats acoustiques doit donc se faire avec prudence.

Consonnes Les consonnes se distinguent des voyelles par la présence d'un resserrement partiel ou complet du conduit vocal en un ou plusieurs endroits. Elles se différencient des voyelles également par un aspect transitoire prédominant, un spectre souvent plus étalé, un niveau d'intensité plus faible et une échelle de temps plus courte. On peut classer les consonnes suivant la présence de vibrations des cordes vocales (consonnes voisées vs non voisées), la qualité de

		F1	F2	F3
voy. fermées	i	308	2064	2976
	y	300	1750	2120
	u	315	764	2027
voy. mi-fermées	e	365	1961	2644
	ø	381	1417	2235
	o	383	793	2283
voy. mi-ouvertes	ɛ	530	1718	2558
	œ	517	1391	2379
	ɔ	531	998	2399
voy.ouv.	a	684	1256	2503

FIGURE 2.14 – Valeurs formantiques moyennes pour les 3 premiers formants des voyelles orales du français d'après [Meunier, 2007].

l'obstruction de l'obstacle (mode d'articulation), et l'endroit où se trouve l'obstruction (lieu d'articulation).

Voisement Les consonnes voisées ou sonores (/b/,/d/,/g/,/v/,/l/,/z/) sont associées à une vibration des cordes vocales. Les consonnes correspondantes non voisées ou sourdes (respectivement /p/,/t/,/k/,/f/,/s/,/ʃ/), ne comportent pas de vibration des cordes vocales.

Modes d'articulation Les modes d'articulations permettent de différencier 3 grandes classes : les occlusives, les constrictives et les nasales.

Les occlusives, ou plosives (/p/,/t/,/k/,/b/,/d/,/g/) impliquent une occlusion complète du conduit vocal, donnant lieu à un silence, suivie d'un relâchement brusque, donnant lieu à un bruit d'explosion. Le silence de l'occlusion se voit sur le spectrogramme, il apparaît comme un vide pour les consonnes sourdes (/p/,/t/,/k/), accompagné d'une barre de voisement pour les occlusives sonores (/b/,/d/,/g/). Le *Voice Onset Time* (VOT) est une mesure de la durée de ces consonnes. C'est la durée entre le moment de l'explosion et le retour du voisement. Il est positif pour les occlusives sourdes (le voisement s'installe après l'explosion) et négatif pour les occlusives sonores (le voisement commence avant l'explosion).

Les constrictives se forment lors d'un resserrement incomplet du conduit vocal. Quand ce resserrement est important cela donne lieu à un écoulement turbulent se traduisant par un bruit de friction, les sons produits sont des fricatives (/f/,/s/,/ʃ/,/v/,/ʒ/,/z/,/ʁ/). Le spectre des fricatives est composé de fréquences plus élevées que les voyelles, allant jusqu'à 20kHz. Lorsque le resserrement est moins important, l'air s'écoule sans bruit de friction, les sons produits sont dit spirantes. Ils sont composés de l'approximante /l/, et des semi-consonnes.

Les nasales (/m/,/n/,/ɲ) se font par occlusion totale de la cavité orale, et ouverture de la cavité nasale. Ce sont des sons voisés comparables aux occlusives sonores (elles sont quelques fois considérées comme telles) à l'exception de l'abaissement du voile du palais. Toutes comme les voyelles nasales, on observe la présence d'anti-formants chez les consonnes nasales, dues au couplage avec la cavité nasale.

Lieu d'articulation Le critère de classification s'appuie sur le point d'articulation (lèvres, dents, palais..) et l'organe articulateur (lèvres, langue). On peut distinguer selon le point d'articulation :

- lèvres : les consonnes labiales (/m/,/b/,/p/)
- dents : les consonnes dentales (/l/,/n/,/t/,/d/)
- lèvres et dents : les consonnes labio-dentales (/f/,/v/)

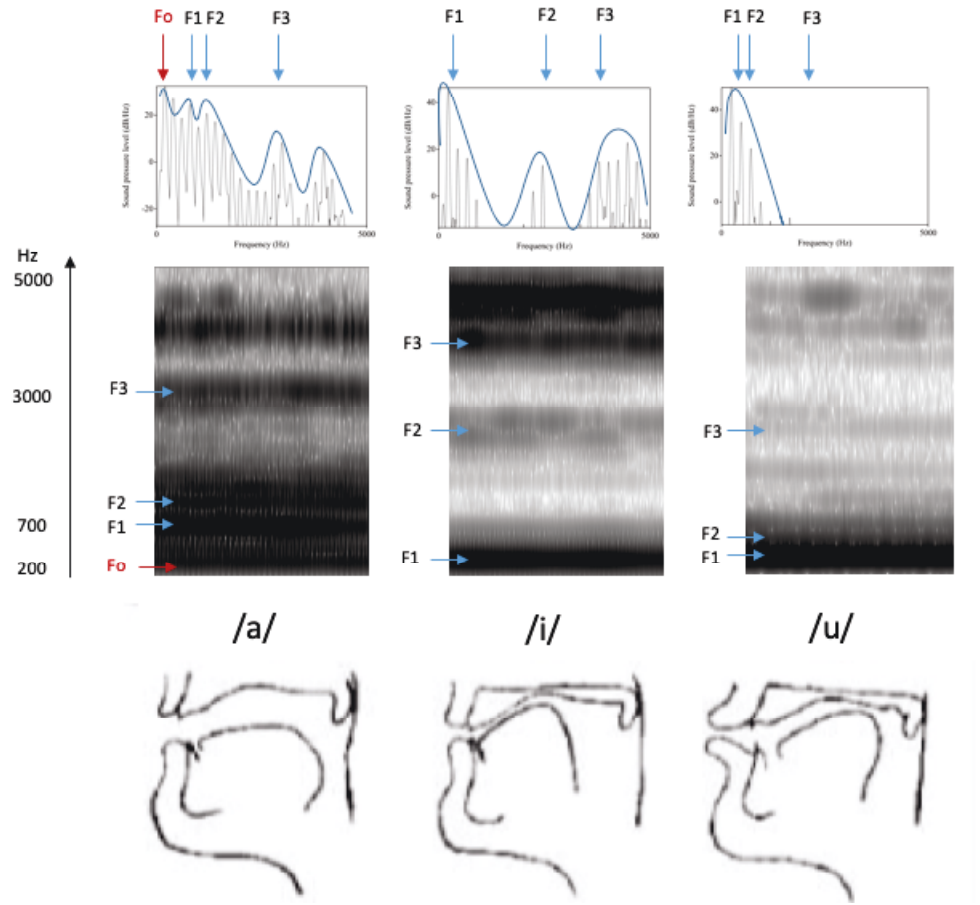


FIGURE 2.15 – Spectrogrammes des voyelles /a/, /i/ et /u/ et leurs spectres correspondants au dessus. Le réglage des spectrogrammes est à bande étroite de manière à faire bien apparaître les formants (ils sont situés au niveau des bandes foncées). Sur les spectres on peut voir la fréquence fondamentale et les harmoniques (raies situées aux multiples de la fréquence fondamentale) ainsi que les formants (pics de l'enveloppe spectrale). Les trois premiers formants F1, F2 et F3 sont indiqués pour les trois voyelles, ainsi que la fréquence fondamentale Fo pour la voyelle /a/. Pour /i/ et /u/, Fo n'est pas assez éloigné de F1 pour qu'on puisse le distinguer. On remarquera pour la voyelle /u/ on distingue difficilement F1 et F2 car ils sont très proches et F3 n'apparaît pas sur le spectre car son intensité est trop faible.

- alvéoles : les consonnes alvéolaires(/s/,/z/) et post-alvéolaire (/ʃ/,/ʒ)
- palais : les consonnes palatales (/j/,/ɥ/,/ɲ)
- voie du palais : les consonnes vélaires(/k/,/g/,/w/)
- luette : les consonne uvulaire (/ʁ/,et /ŋ/ emprunt de l'anglais)
- glotte : les consonne glottales (/h/ emprunt de l'anglais)

Semi-voyelles ou semi-consonnes Les semi-voyelles ou semi-consonnes, aussi appelées glissantes, sont une classe intermédiaire entre les voyelles et les consonnes. Elles présentent la même articulation que les voyelles mais ont une place dans la syllabe similaire aux consonnes. Elles sont au nombre de trois en français : /j/,/w/,/ɥ/. Elles présentent une structure formantique instable, à la différence des voyelles, due à un resserrement plus important des articulateurs et par une articulation en mouvement.

2.3 Modifications de la voix dans MP

Les études sur la voix dans la maladie de Parkinson parlent souvent de dysarthrie hypokinétique (qui signifie réduction de l'amplitude des mouvements des muscles responsables de l'articulation) pour catégoriser les troubles de la voix des parkinsoniens. Les différentes composantes de la parole affectées par la dysarthrie parkinsonienne sont :

- la **prosodie** : une perte des modulations d'intensité et de hauteur donne à la voix un caractère monotone, le débit est altéré et on constate aussi des troubles de la fluence (palilalies, bredouillements ...) ;
- l'**articulation** : la précision articulatoire des voyelles et des consonnes est altérée ;
- la **phonation** : l'intensité de la voix diminue (le patient devient hypophone), la hauteur moyenne s'abaisse ou s'élève, la hauteur et l'intensité deviennent instables, et le timbre devient soufflé, voilé, éraillé ;
- le **rythme** : la capacité à maintenir un rythme de parole constant s'altère.

Dans la suite nous détaillerons les dysfonctionnements de la voix que l'on trouve dès le début de la MP.

2.3.1 Prosodie

L'insuffisance prosodique constituerait la marque la plus spécifique des troubles de la parole dans la maladie de Parkinson [Viallet and Teston, 2007]. Elle se caractérise chez les sujets MP débutants par une monotonie de la mélodie (diminution de la variation de la fréquence fondamentale F0), par une monotonie de l'intensité, et par une diminution du nombre de pauses de plus de 60 ms. La dysprosodie serait le résultat d'une diminution de l'amplitude du mouvement du larynx et des muscles respiratoires causés par une rigidité excessive [Rusz et al., 2013b]. Une étude acoustique en 2011 a montré que les problèmes de prosodie étaient présents chez plus de 60% des 23 MP débutants non traités testés [Rusz et al., 2011a]. Cette même étude montre que le monologue et la lecture de phrases émotionnelles mettent plus en avant la diminution de la variation de la fréquence fondamentale que la lecture de texte. Une étude plus récente sur 24 MP débutants non traités a même obtenu un taux de réussite de 81,3 % de classification (MP débutants vs sains), en analysant juste la variation de la fréquence fondamentale pendant un monologue [Rusz et al., 2011b].

2.3.2 Articulation

Un déficit d'articulation a été mis en évidence dans la maladie de Parkinson par de nombreuses études [Rusz et al., 2011a, Rusz et al., 2011b, Skodda et al., 2012], et son analyse acoustique permettrait à elle seule de discriminer des parkinsoniens débutants de sujets sains avec un taux de réussite de plus de 88% (d'après une étude acoustique qui a porté sur 24 MP débutants non traités et 22 sujets sains [Novotný et al., 2014]). Les problèmes d'articulation se voient à la fois dans l'articulation des voyelles et dans l'articulation des consonnes, et se manifestent par une diminution des contrastes acoustiques.

Articulation des voyelles : On note chez les parkinsoniens une tendance à la dédifférenciation des voyelles. Les formants ayant une fréquence naturellement élevée voient leur fréquence diminuer, ce qui est le cas du 2^e formant de la voyelle /i/ et du 1^{er} formant de la voyelle /a/. Les formants ayant normalement une fréquence basse subissent une augmentation de leur fréquence, ce qui est le cas du 2^e formant de la voyelle phonétique /u/. Ceci a pour conséquence une diminution de la surface du triangle vocalique (VSA, pour *Vowel Space Area*) et de l'index d'articulation

vocalique (VAI) [Skodda et al., 2012]. Cette centralisation des formants traduirait une diminution de l'amplitude des mouvements de la langue et des lèvres ; ce serait le corolaire vocal de la bradykinésie (réduction de la vitesse et de l'amplitude des mouvements) [Rusz et al., 2013b]. Elle apparaîtrait dès le début de la maladie (chez des patients débutants et non traités) [Rusz et al., 2013a]. D'après ces auteurs, la centralisation des formants a lieu surtout lors de la parole spontanée. On la trouve de façon plus atténuée pendant la lecture d'un texte, et n'a pas du tout lieu lors de prononciation de voyelles soutenues. Les auteurs de cette étude l'expliquent par le fait que pendant la lecture de texte, le patient peut se concentrer sur l'articulation, alors que pendant le discours spontané il doit d'abord se concentrer sur le sens de ce qu'il est en train de dire et donc prête moins d'attention à l'articulation.

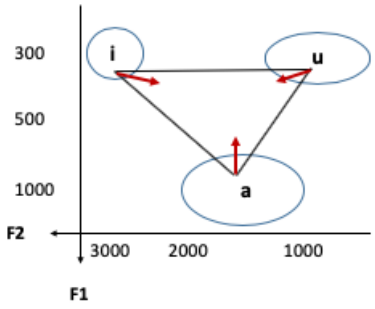


FIGURE 2.16 – Représentation des premier formant F1 et deuxième formant F2 des voyelles cardinales du triangle vocalique chez les sujets sains. Les flèches rouges indiquent la modification de ces formants chez les sujets atteints de la maladie de Parkinson, entrainant une diminution de la surface du triangle vocalique.

Articulation des consonnes : L'articulation des consonnes s'effectue de manière imprécise chez les patients parkinsoniens, et ce également chez des MP débutants non traités. Les patients ont tendance à ne pas fermer complètement leur conduit vocal lors de la prononciation de consonnes occlusives orales (p, t, k, b, d, g). Cela crée une fuite d'air turbulent qui peut se détecter à la place du silence qui est censé avoir lieu pendant l'occlusion. Les consonnes occlusives orales ressemblent alors un peu plus à des fricatives (f, s) et représentent l'anomalie articulatoire la plus marquée dans la dysarthrie parkinsonienne [Locco, 2005, Pinto et al., 2010]. Le rapport signal sur bruit permet de mesurer cet effet. Il correspond à $10log(Is/In)$ avec Is l'intensité acoustique qu'on peut mesurer pendant le son voisé et In celle qui correspond à la partie censée être silencieuse pendant l'occlusion [Novotný et al., 2014].

Un autre défaut de l'articulation provient d'une mauvaise coordination entre les muscles laryngés (cordes vocales) et supralaryngés (langue, lèvres, mandibules), ce qui entraîne une articulation imprécise. Les études [Rusz et al., 2015, Novotný et al., 2014] ont montré que la durée des consonnes occlusives non voisées (p,t,k), mesurée grâce au paramètre *Voice Onset Time* (VOT), était alors augmentée chez les sujets MP débutants.

Ces études ont aussi montré des anomalies formantiques chez les patients parkinsoniens débutants, expliquées par des perturbations dans les mouvements de la langue lors de tâches de diadococinésie (DDK) [Novotný et al., 2014].
Les différents problèmes d'articulation rencontrés ont également comme impact une diminution du débit de parole lors des tâches DDK. On l'observe notamment chez les parkinsoniens débutants non traités [Novotný et al., 2014].

2.3.3 Phonation

Les problèmes de phonation dans la maladie de Parkinson concernent la hauteur, l'intensité et le timbre. La tâche vocale qui les met le mieux en évidence est la prononciation de voyelles soutenues : on demande aux sujets de prononcer la voyelle /a/ le plus longtemps possible sans respirer. Lors de ce type de tâche, les patients parkinsoniens, même ceux récemment diagnostiqués, ont du mal à maintenir la hauteur de la voix constante, et cela ne fait que s'accentuer avec la progression de la maladie. L'instabilité à moyen terme de la hauteur se mesure par l'écart type de la fréquence fondamentale (Fo SD). L'instabilité à court terme de la fréquence fondamentale se traduit par des variations de fréquence entre chaque cycle d'oscillation (appelées *jitter*).

Les patients atteints de la maladie de Parkinson souffrent également d'une réduction de l'intensité moyenne et de sa stabilité. L'instabilité de l'intensité sur le moyen terme a été mise en évidence chez des MP débutants non traités lors de tâche de diadococinésie, où on a observé une augmentation du *Relative Intensity Range Variation* (RIRV) [Rusz et al., 2011a]. L'instabilité de l'intensité sur le court terme, apparaît surtout lors de voyelles soutenues et se traduit par une variation de l'amplitude entre chaque cycle d'oscillation (dénommée *shimmer*) [Rusz et al., 2011a].

Le timbre de la voix des parkinsoniens est aussi altéré et apparaît comme légèrement soufflé et éraillé. Il serait dû à un accolement incomplet des cordes vocales, qui a été mis en évidence par des analyses laryngoscopiques [Jiang et al., 1999]. Il peut être mesuré par le paramètre *Harmonic-to-Noise ratio* (HNR). Ce paramètre indique l'amplitude du bruit par rapport aux composantes tonales. Il est plus élevé chez les parkinsoniens que chez les sujets sains et ce dès les premières années après le diagnostic [Rusz et al., 2011a].

Une autre composante de la parole qui peut influencer la phonation est la respiration. Les parkinsoniens ont des problèmes de respiration (ils prennent des inspirations moins profondes et ont du mal à coordonner respiration et parole) qui font que l'intensité de leur voix est plus faible, surtout quand la maladie est un peu plus avancée [Countryman et al., 2003]. Cela a aussi comme conséquence de diminuer la durée maximale de phonation (MPT, pour *Maximum Phonation Time*) des femmes parkinsoniennes débutantes, lorsqu'elles doivent dire des voyelles soutenues le plus longtemps possible [Huh et al., 2015]. Il est intéressant de noter que cela concerne les femmes mais pas les hommes. En effet certains paramètres acoustiques (Fo, MPT, les coefficients cepstraux, VSA...) et leur évolution au cours de la maladie, sont très sensibles au genre [Skodda et al., 2012,Tsanas et al., 2011]. Donc pour ce type de paramètres, il est préférable de faire des analyses séparées pour les hommes et les femmes.

2.3.4 Rythme

Effectuer des mouvements automatiques à un rythme stable est quelque chose qui est connu pour être difficile dans la maladie de Parkinson. Cette instabilité motrice serait la conséquence d'un dysfonctionnement des ganglions de la base qui ne pourraient plus assurer correctement la préparation et le maintien de séquences motrices simples qui s'effectuent normalement de manière quasi automatique [Iansek et al., 1995]. Cette difficulté apparaît notamment dans la parole : les parkinsoniens ont du mal à répéter une série de syllabes (/pa/ par exemple) à un rythme lent et régulier. Cette difficulté étant accrue quand le rythme leur est imposé et encore plus quand il s'agit d'alterner entre deux syllabes différentes (/pa/,/ti/) [Skodda et al., 2013]. On retrouve cette difficulté dès les premières années après le diagnostic. En effet une étude sur 50 MP débutants traités et 32 sujets sains a montré que le coefficient de variation relative du rythme est significativement plus élevé chez les parkinsoniens que chez les sujets sains, lors de la répétition de syllabes à un rythme choisi et imposé [Skodda, 2015]. Les auteurs de cette étude ont aussi montré une corrélation entre le score UPDRS (caractérisant l'avancement de

la maladie) et le nombre maximal de syllabes que les sujets pouvaient dire par seconde, quand on leur demandait de répéter la syllabe /pa/ le plus rapidement possible. Cette corrélation n'a cependant pas été trouvée avec la variation relative du rythme. Les auteurs en ont conclu que la vitesse de répétition et sa régularité correspondaient à des domaines différents de performance motrices basiques, avec possiblement des physiopathologies différentes.

2.3.5 Effet des traitements pour la maladie de Parkinson sur la voix

Certains troubles de la voix dus à la maladie de Parkinson s'améliorent avec des traitements, et ce même chez les parkinsoniens débutants. Nous allons d'abord nous intéresser à un traitement comportemental : le *Lee Silverman Voice Training* (LSVT) dont le but est exclusivement d'améliorer les problèmes de voix dus à la maladie de Parkinson. Ensuite nous verrons quelles sont les conséquences des traitements pharmaceutiques dopaminergiques, que l'on donne pour améliorer les dysfonctionnements moteurs de la maladie de Parkinson, sur la voix.

2.3.5.1 LSVT (*Lee Silverman Voice Training*)

La LSVT est une technique d'orthophonie utilisée depuis 2004 dont le but est de limiter la diminution d'intensité vocale et la perte de prosodie, en améliorant l'accolement des cordes vocales et en renforçant de façon générale l'activation des muscles laryngés et leur contrôle. L'entraînement se déroule pendant 16 sessions d'1h réparties de façon homogène sur 4 semaines. Durant ces sessions le patient est invité à prononcer avec une voix forte des voyelles soutenues, en faisant varier ou pas la hauteur de la voix, et à parler d'une voix forte en se concentrant sur l'intensité de sa voix. D'une manière générale on conseille au patient de "penser fort" ("*think loud*"), pour améliorer le traitement de l'information sensorielle auditive d'origine proprioceptive. En effet le patient parkinsonien hypophonique a tendance à ne pas se rendre compte qu'il ne parle pas assez fort [Viallet and Teston, 2007]. Les effets bénéfiques sur l'intensité de la voix et la prosodie apparaissent généralement au bout d'un mois et sont encore visibles 2 ans après [Ramig et al., 2001].

2.3.5.2 Traitements dopaminergiques

L'effet des traitements pharmaceutiques dopaminergiques sur la voix des patients a été mis en évidence récemment sur un groupe de 19 patients MP débutants [Rusz et al., 2013b]. Les traitements dopaminergiques ont induit des améliorations, classées de la plus à la moins importante, dans les domaines suivants (se référer au Tableau 2.2 pour la signification des paramètres acoustiques) :
- intensité de la voix (Int SD pour monologue et lecture) ;
- qualité de la voix (*jitter*, *shimmer*, HNR, RPDE, PPE) ;
- intonation (F0 SD pour monologue et lecture) ;
- articulation des voyelles (VAI, F2i/F2u).

Pour ces patients au stade débutant, les améliorations sont visibles dans les analyses acoustiques mais n'apparaissent pas dans les analyses perceptives (item 18 de l'UPDRS inchangé). La dopamine semble donc avoir un impact sur l'intensité, la qualité et l'intonation de la voix, mais d'après une autre étude, elle n'aurait pas d'influence sur la régularité du débit de la parole, le nombre de pauses et le rythme [Skodda, 2015].

Les traitements orthophoniques (de type LSVT) et dopaminergiques ont un impact positif sur certains troubles vocaux rencontrés dans la maladie de Parkinson. L'influence de ces traitements sur la voix doit donc être prise en compte lors de l'interprétation d'analyses vocales chez des patients parkinsoniens traités.

2.3.6 Particularités de la voix au stade préclinique de la maladie de Parkinson

Nous avons pu voir que plusieurs études avaient montré qu'il était possible de détecter la maladie de Parkinson chez des parkinsoniens débutants en analysant simplement la voix. Mais l'enjeu réside surtout dans le fait de pouvoir diagnostiquer plus tôt la maladie qu'il n'est possible à l'heure actuelle avec l'examen moteur. Quelques équipes ont donc cherché à savoir si certains troubles de la voix n'apparaîtraient pas avant les symptômes moteurs qui servent au diagnostic actuel, et pourraient ainsi par la suite servir de marqueurs de diagnostic très précoce.

2.3.6.1 Etude rétrospective à partir d'extraits télévisés

En 2004 une étude a pour la première fois mis en évidence des changements mesurables dans la voix durant le stade préclinique d'un individu atteint de la maladie de Parkinson. Cet individu donnait régulièrement des interviews et des conférences à la télévision. En analysant les enregistrements vidéo qui dataient de 7 ans avant le diagnostic jusqu'à 3 ans après celui-ci, et en les comparant avec des enregistrements d'un sujet sain apparié, les auteurs ont montré que les variations de la fréquence fondamentale commençaient à diminuer significativement à partir de 5 ans avant le diagnostic [Harel et al., 2004]. Cette étude a le mérite d'être la première étude longitudinale à effectuer une analyse acoustique de la voix d'un patient parkinsonien pendant sa phase prodromique. Néanmoins il faudrait refaire cette analyse sur un nombre plus important de sujets pour pouvoir valider ces résultats.

2.3.6.2 Etudes sur les RBD (*REM sleep Behaviour Disorder*)

Pendant la phase de sommeil paradoxal, on a normalement une atonie : nos mouvements sont inhibés. Certaines personnes n'ont pas cette atonie, on nomme ce dysfonctionnement RBD pour *REM (Rapide Eye Mouvement) sleep Behaviour Disorder*. Deux tiers des individus atteints de la maladie de Parkinson souffrent aussi de RBD. Inversement quasiment toutes les personnes ayant un RBD vont développer un syndrome parkinsonien. En effet au bout de 14 ans, 91% des patients RBD ont développé un syndrome parkinsonien [Iranzo et al., 2014]. Parmi les syndromes parkinsoniens développés par les RBD on trouve la maladie de Parkinson et d'autre maladies proches qui, en plus des symptômes parkinsoniens courants, comprennent d'autres troubles (le plus courant étant la démence) : il y a notamment la démence à corps de Lévy (DCL), et plus rarement l'atrophie multisystématisée (MSA). Les RBD qui n'ont pas encore développé de syndrome parkinsonien peuvent donc être considérés comme étant dans la phase prodromique d'un syndrome parkinsonien. L'analyse de leur voix peut alors donner des indications sur les marqueurs vocaux prédictifs de la maladie de Parkinson.

Une étude longitudinale a montré qu'en effectuant une analyse perceptive de la voix de 78 RBD tous les ans jusqu'au diagnostic d'un syndrome parkinsonien, on pouvait estimer le début des perturbations vocales à 7 ans avant le diagnostic pour ceux qui ont finalement développé la maladie de Parkinson, et à 15 ans avant le diagnostic pour ceux qui ont développé une DCL [Postuma et al., 2012].

Une étude acoustique a quant à elle analysé quels paramètres acoustiques différaient significativement chez 16 RBD en phase prodromique par rapport à des sujets sains [Rusz et al., 2015]. Les auteurs ont testé des paramètres en rapport avec la phonation, l'articulation, et la prosodie. Ils ont trouvé que l'articulation était le domaine le plus affecté, suivi de la phonation, puis de la prosodie. Parmi les paramètres acoustiques discriminants on note une irrégularité du débit (DDK reg) lors des tâches de diadococinésie, une diminution de l'énergie spectrale (RFA) lors du monologue ainsi qu'une augmentation de disfluences, et une apériodicité phonatoire (DUV) lors des voyelles soutenues (cf. Tableau 3.2). Il faut faire attention à ne pas interpréter ces paramètres comme étant forcément des marqueurs de prédiction de la maladie de Parkinson

car les RBD pourront développer un autre syndrome parkinsonien, comme la démence à corps de Lévy ou la MSA. Or ces maladies sont accompagnées de perturbations vocales qui peuvent être légèrement différentes de celles que l'on trouve dans la maladie de Parkinson [Huh et al., 2015,Müller J et al., 2001]. Une étude complémentaire s'est focalisée sur la comparaison entre ces RBD en phase prodromique et des MP débutants [Rusz et al., 2015]. Les auteurs ont noté qu'en moyenne les RBD étaient moins affectés vocalement que les parkinsoniens débutants, surtout en ce qui concerne la prosodie. Les paramètres F0 SD (variation de la fréquence fondamentale) et NoP (nombre de pauses) pour le monologue et le VOT pour la DDK tâche sont les paramètres les plus discriminants quand on compare les RBD avec les parkinsoniens débutants.

Paramètre	Description	Tâche
Phonation		
HNR	Harmonics-to-Noise Ratio : Amplitude des composantes tonales par rapport au bruit	voyelle soutenue
MPT	Maximum Phonation Time : durée maximale de phonation	voyelle soutenue
F0 SD	Standard Deviation (écart type) de la fréquence fondamentale (F0)	voyelle soutenue
DUV	Degree of Unvoiced Segment : fraction des segments silencieux ($< 0,45$)	voyelle soutenue
jitter	Variabilité de F0 d'un cycle à l'autre	voyelle soutenue
shimmer	Variation relative de l'amplitude entre 2 cycles consécutifs	voyelle soutenue
Articulation		
VOT	Voice Onset Time : durée d'une consonne occlusive	DDK
DDK rate	Taux diadococinésie (DDK) : nombre de syllabes par seconde	DDK
DDK reg	Régularité DDK : écart type des distances entre 2 centres syllabiques consécutifs	DDK
RFA	Resonant Frequency Attenuation	DDK
SNR	Signal-to-Noise Ratio	DDK
1FT	First Formant Trend : régression de la fréquence du 1er formant	DDK
2FT	Second Formant Trend : régression de la fréquence du 2^e formant	DDK
VSQ	Vowel Similarity Quotient : autocorrélation de la voyelle sur sa durée totale	DDK
VSQ30	VSQ sur une durée de 30ms à partir du début de la voyelle	DDK
VVQ	Vowel Variability Quotient : variabilité dans les longueurs des voyelles	DDK
CST	Consonent Spectral Trend : régression du spectre de la consonne	DDK
RIRV	Relative Intensity Range Variation : variation relative de l'intensité	DDK
RRIS	Robust Relative Intensity Slope	DDK
SDCV	Spectral Distance Change Variation	DDK
RFPC	Robust Formant Periodicity Correlation	DDK
VSA	Vowel Space Area : aire du triangle vocalique	monologue
VAI	Vowel Articulation Index = (F1a+F2i)/(F1i+F1u+F2a+F2u)	monologue
F2i/F2u	Rapport des 2^e formants des voyelles /i/ et /u/	monologue
Prosodie		
F0 SD	Standart Deviation de F0 : variation de la hauteur (intonation)	monologue
Int SD	Standart Deviation de l'Intensité après suppression des silences > 60ms	monologue
NoP	Nb de Pauses par rapport au temps parlé après suppression des silences < 60 ms	monologue
PDW	Percentage of Disfluent Words : nombre de disfluences sur nombre total de mots	monologue
rythm	Mesure de la capacité à reproduire les rythmes d'une lecture après écoute	lecture rythmée
Rythme		
COV	Coefficient of Variation : variation relative du rythme	répét lente de syll.
pa-ti ratio	Intervalle pa-ti par rapport à intervalle pa-pa	répét lente de syll.

TABLE 2.2 – Tableau regroupant les paramètres acoustiques les plus discriminants pour le diagnostic précoce de la maladie de Parkinson d'après la littérature, et les meilleures tâches vocales qui permettent de les extraire. DDK : tâche de diadococinésie.

Maintenant que nous avons décrit les différentes modifications de la voix dans la maladie de Parkinson, nous allons voir comment les études de la littérature les utilisent pour construire des modèles de classification afin de détecter MP.

Chapitre 3

Etat de l'art : Classification MP vs sain par l'analyse acoustique de la voix

Dans ce chapitre nous présentons les différentes méthodes utilisées dans la littérature pour détecter MP par la voix. Nous commencerons par un rappel du fonctionnement des principales méthodes de classification et leur évaluation. Puis nous présenterons un état de l'art des classifieurs utilisés et des performances obtenues dans le cadre de la détection de MP.

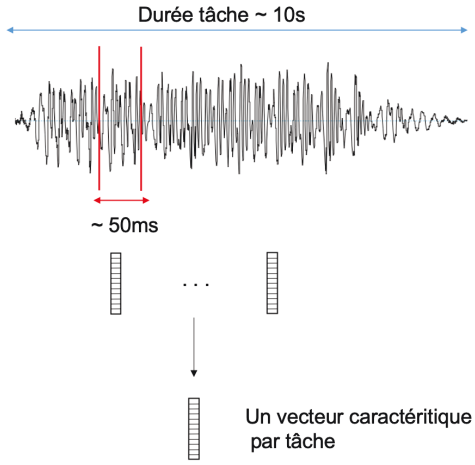
3.1 Méthodes de classification

La classification de la voix, que ce soit pour identifier un mot, un locuteur, une particularité (langue, accent, genre, émotion ...) ou une pathologie vocale, se décompose en plusieurs phases. La première phase est celle de l'extraction de paramètres vocaux à partir du signal brut. Ces paramètres peuvent subir une étape de traitement, visant à réduire le bruit ou les distorsions dues au canal d'enregistrement. Ils peuvent également faire l'objet d'une sélection, visant à supprimer les paramètres non pertinents ou redondants, avec au final un ou plusieurs vecteurs caractéristiques par sujet. Après cette étape s'ensuit la phase de construction d'un modèle de classement automatique, faisant généralement appel à l'apprentissage automatique supervisé. Cette phase commence par l'entraînement du modèle de classification à partir des vecteurs caractéristiques labellisés (dont la classe est précisée) extraits de données servant à l'apprentissage. Ce modèle peut soit modéliser les caractéristiques d'une classe, dans ce cas il est dit génératif, soit modéliser la "limite" entre deux classes, il est alors dit discriminant. Après l'étape d'apprentissage du modèle, a lieu une étape de test pendant laquelle de nouvelles données sont classées. Les hyperparamètres du modèle peuvent être ajustés de manière à optimiser le taux de données test bien classées. Pour finir, une étape de validation où de nouvelles données sont testées permet de généraliser la performance du classifieur. Ces différentes étapes sont détaillées ci-dessous.

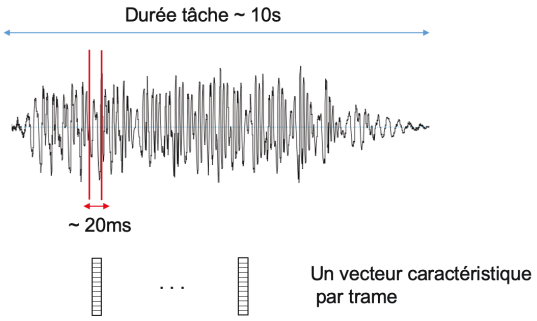
3.1.1 Extraction de paramètres

La première étape d'une classification est l'extraction de paramètres. Ces paramètres peuvent être regroupés en deux grands groupes. D'un côté les paramètres globaux, (long-terme), et de l'autre les paramètres dits niveau trame (court-terme). Nous appelons paramètre global, un paramètre représentant la tâche entière, cela peut être un dénombrement de certains événements (nombre de pauses, de dysfluences) qui ont eu lieu durant une tâche vocale, ou une moyenne de paramètres locaux (comme Fo, jitter, shimmer..), calculés généralement sur des fenêtres temporelles de l'ordre de 50ms. Ce type d'extraction aboutit à un vecteur de paramètres par sujet et par tâche, cf. Figure 3.1a.

Le deuxième type de paramètres sont les paramètres cout-terme. Ils sont extraits sur des fenêtres temporelles de l'ordre de 20ms, souvent toutes les 10ms, et sont considérés ensemble pour décrire le sujet et la tâche en question (sans être moyennés), cf. Figure 3.1b. Les modèles sont généralement construits pour décrire la distribution de ces paramètres, et les tests se font trame par trame. A la différence des paramètres globaux, construits la plupart du temps à partir de paramètres locaux moyennés, pour les paramètres court-terme, ce sont les scores de classification issus de chaque trame qui sont moyennés.



(a) Analyse globale



(b) Analyse sur temps courts

FIGURE 3.1 – Deux approches pour la classification à partir de l'analyse de la voix. En a) la méthode d'analyse globale, ou sur temps long, la classification se fait à l'échelle de la tâche. En b) la méthode d'analyse sur temps courts, la classification s'effectue à l'échelle de la trame.

L'étape d'extraction des paramètres est souvent accompagnée de traitements dans le but de supprimer les bruits, les distorsions et les silences.

3.1.2 Modèles de classification

Considérons que l'on souhaite classer les voix en 2 classes (Y) à partir des paramètres vocaux (X). Les modèles génératifs vont apprendre à décrire les densités de probabilité conditionnelles $p(X|Y)$, alors que les modèles discriminants apprennent la probabilité conditionnelle $p(Y|X)$. Par extension on considérera également comme discriminant les classifieurs qui n'apprennent aucune distribution de probabilité. Des exemples d'algorithmes classiques pour ces deux types de modèles sont détaillés ci-dessous. Leur adaptation et leurs variantes utilisées dans le contexte spécifique de reconnaissance du locuteur seront présentées dans la partie 3.3.1.

3.1.2.1 Modèles génératifs

Classifieur de distance minimale La façon la plus simple de modéliser la distribution des probabilités conditionnelles $p(X|Y)$, soit les caractéristiques de chaque classe, est de moyenner les vecteurs caractéristiques X, issues des données d'entraînement, pour chaque classe. Un vecteur caractéristique moyen est alors représentatif de chaque classe. Pour chaque sujet test, les distances (Euclidiennes par exemple) entre son vecteur caractéristique (moyenné si besoin sur toutes les trames) et les vecteurs caractéristiques moyens représentatifs des deux classes sont calculées. La différence ou le rapport entre ces distances est le score de classification. Le seuil de classification est généralement 0 pour la différence et 1 pour le ratio, de telle manière que la classe attribuée au sujet test corresponde à la classe du vecteur caractéristique moyen le plus proche.

Quantification vectorielle La modélisation de la distribution des caractéristiques de chaque classe par leur moyenne, est une méthode simple mais peu précise. Or décrire $p(X|Y)$ plus précisément peut s'avérer très important pour la classification (surtout lorsque les distributions sont proches et non gaussiennes). La quantification vectorielle (VQ), utilisée à l'origine pour la compression de données, est une technique de quantification qui permet de décrire une distribution par un vecteur de dimension plus petite. L'étape d'apprentissage est une étape de partitionnement, utilisant des algorithmes de clustering, comme le partitionnement en k-moyennes. K points (le nombre K étant choisi par l'utilisateur) dit centroïdes sont placés au hasard dans l'espace des observations. A chaque donnée X d'entraînement est associée le centroïde le plus proche (d'après un calcul de distance euclidienne par exemple). L'ensemble des données X associées à un même centroïde forme un cluster. Les centroïdes sont ensuite déplacés au barycentre des points de leur cluster. Les données X sont réassignées au nouveau centroïde le plus proche, etc... jusqu'à convergence. Au final les données X sont divisées en K cluster, d'environ le même nombre de points, minimisant une fonction de coût (par exemple la somme des carrés des distances entre les points X et les centroïdes associés). La distribution de X pour chaque classe est alors modélisée par le vecteur formé par les k centroïdes finaux, appelé aussi codebook.

Dans le cas où il n'y a qu'un seul vecteur caractéristique X par sujet test (analyse globale), l'étape de test consiste à calculer la distance entre son vecteur caractéristique et le centroïde le plus proche, et ce pour chaque classe. Dans le cas de l'analyse court terme, la moyenne de la distance entre les vecteurs caractéristiques et le centroïde le plus proche est calculée sur l'ensemble des trames, pour chaque classe, résultant en une valeur de "distorsion" par classe. Tout comme pour le classifieur de distance minimale, on choisit généralement d'attribuer au sujet test la classe correspondant à la distorsion la plus faible.

Modèle de Mélanges Gaussiens, GMM La quantification vectorielle, lors du clustering de l'apprentissage, attribue une seule classe aux points X, on peut considérer cette décision comme "dure". Une alternative, qui permettrait de mieux représenter les enregistrements à grande variabilité acoustique, serait de prendre une décision dite "souple" en introduisant un modèle probabiliste. Une fonction de densité de probabilité multidimensionnelle des vecteurs caractéristiques décrirait alors la distribution des vecteurs caractéristiques (au lieu d'utiliser un vecteur de centroïde comme dans la VQ). Les différentes classes seraient alors décrites par les paramètres de ces fonctions de densité de probabilité, et l'apprentissage serait basé sur des décisions souples prenant en compte les probabilités d'appartenance de X à chaque classe. Plusieurs fonctions peuvent modéliser la densité de probabilité des vecteurs caractéristiques, nous nous intéresserons à la fonction la plus utilisée qui est celle du modèle des mélanges de Gaussiennes (GMM).

Cas du modèle mono-gaussien : La distribution de n vecteurs caractéristiques $\mathbf{x}$ de dimension d peut être modélisée par une unique distribution gaussienne (ou normale) de vecteur moyen $\boldsymbol{\mu}$ et de matrice de covariance Σ de dimension $d \times d$. La densité de probabilité de cette distribution gaussienne est donnée par :

$$g(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}(det\Sigma)^{1/2}} e^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{3.1}$$

Ce qui donne dans le cas à une dimension :

$$g(x) = \frac{1}{(\sqrt{2\pi})(\sigma)} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \qquad \text{avec } \sigma \text{ l'écart type} \tag{3.2}$$

Dans le cas multidimensionnel, si les coefficients de $\boldsymbol{x}$ sont indépendants et de même variance (Σ est proportionnelle à la matrice identité), la densité de probabilité aura une allure sphérique (cf. Figure 3.2). Si les coefficients sont indépendants mais de variances différentes (Σ est une

matrice diagonale, avec des valeurs non égales sur la diagonale), la fonction de densité de probabilité aura une allure elliptique le long d'axes parallèles aux axes principaux. Si les coefficients sont corrélés (Σ est une matrice non diagonale, dite pleine) la fonction de densité de probabilité aura une allure elliptique le long d'axes non parallèles aux axes principaux. On peut noter que si les coefficients sont indépendants, les projections de la fonction gaussienne multidimensionnelle sur les axes principaux suffisent pour la reconstituer (ce qui n'est pas le cas si les coefficients sont corrélés entre eux).

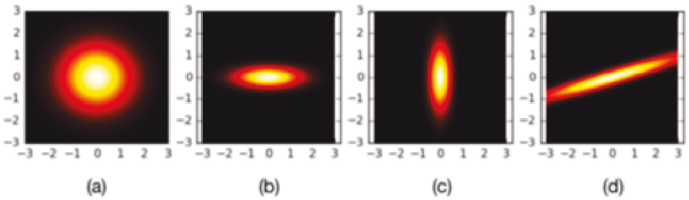


FIGURE 3.2 – Distributions gaussiennes en dimension deux avec une matrice de covariance proportionnelle à la matrice identité (a), diagonale (b et c) et pleine (d). Source : [Larcher, 2018]

Les paramètres μ et Σ définissant la distribution gaussienne qui se rapproche le plus des n observations sont estimés d'après le critère de maximum de vraisemblance de la façon suivante. Le vecteur μ est alors égal à la moyenne des n $\boldsymbol{x}$ et Σ égal à la matrice de covariance de ces n vecteurs. L'inconvénient du modèle mono-gaussien est qu'il ne permet de modéliser correctement que les distributions monomodales (un seul pic dans la distribution).

Modèle des mélanges de Gaussiennes : De manière à représenter plus précisément les distributions plus complexes, on n'utilise plus seulement une gaussienne mais une combinaison convexe de K gaussiennes, K étant fixé par l'utilisateur.

La densité de probabilité est alors décrite par :

$$\begin{cases} p(x) = \sum_{k=1}^{K} \alpha_k g_k(x) \\ avec \quad \sum_{k=1}^{k} \alpha_k = 1 \quad et \quad \forall k \quad \alpha_k > 0 \end{cases} \tag{3.3}$$

g_k étant la kème gaussienne (définie par μ_k et Σ_k) et α_k son poids, au sein de la somme pondérée.

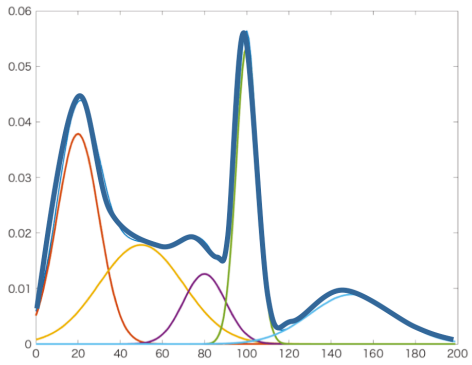


FIGURE 3.3 – Exemple de modèle de mélange gaussien (GMM) en dimension 1. La densité de probabilité de ce GMM est représentée en traits épais, les différentes gaussiennes qui le composent en traits fins.

Les paramètres α_k, $\boldsymbol{\mu_k}$, et Σ_k du mélange de gaussiennes décrivant au mieux les données sont estimés à l'aide d'un algorithme espérance-maximisation (EM).

EM est composé des étapes suivantes :

-**Initialisation** : les paramètres $\boldsymbol{\mu_k}$, et Σ_k sont initialisés au hasard pour chacune des gaussiennes. Les α_k initiaux sont généralement fixés à 1/K.

-**Estimation** : à chaque vecteur caractéristique $\boldsymbol{x_i}$ (i=1..n) est associée la probabilité *a posteriori* qu'il soit généré par la gaussienne d'indice k, (k=1..K), calculée par la règle d'inversion de Bayes. On définit le paramètre de pondération a_{ki} par la valeur de cette probabilité.

$$a_{ki} = p(k|\boldsymbol{x_i}) = \frac{p(\boldsymbol{x_i}|k)p(k)}{\sum_{k=1}^{K} p(\boldsymbol{x_i}|k)p(k)} \tag{3.4}$$

Les p(k) sont les probabilités *a priori* des gaussiennes, on les suppose égales. p(x—k) est la fonction de probabilité décrivant la kème gaussienne ($p(x|k) = g_k(x)$).

-**Maximisation** : Pour chaque gaussienne k on modifie les paramètres $\boldsymbol{\mu_k}$, et Σ_k, de telle manière que $\boldsymbol{\mu_k}$ soit égale à la moyenne des $\boldsymbol{x_i}$ pondérée par les coefficients de pondération a_{ki}, et que Σ_k soit égale à la covariance des $\boldsymbol{x_j}$ pondérée par les mêmes coefficients.

$$\boldsymbol{\mu_k} = \frac{\sum_{i=1}^{n} a_{ki}\boldsymbol{x_i}}{\sum_{i=1}^{n} a_{ki}} \tag{3.5}$$

$$\Sigma_k = \frac{\sum_{i=1}^{n} a_{ki}(\boldsymbol{x_i} - \boldsymbol{\mu_k})^t(\boldsymbol{x_i} - \boldsymbol{\mu_k})}{\sum_{i=1}^{n} a_{ki}} \tag{3.6}$$

Les poids des gaussiennes α_k sont aussi ajustés à chaque itération comme étant la moyenne des a_{ki}.

Après cette étape de maximisation, la vraisemblance $P(X|\lambda)$ des données par rapport au modèle formé par le mélange des nouvelles gaussiennes (soit la probabilité d'avoir le jeu de données X sachant qu'on a ce modèle de densité de probabilité (noté λ)) a augmenté.

-**Réitération** de l'étape d'estimation et de maximisation jusqu'à convergence de la vraisemblance.

Si on n'ajuste pas les variances et qu'on n'attribue qu'une gaussienne par vecteur caractéristique et non une pondération de toutes les gaussiennes, on retombe sur l'apprentissage K-moyennes de la quantification vectorielle.

Ainsi les GMM permettent de décrire plus précisément que la VQ la distribution des données $\boldsymbol{x_i}$, en estimant le mélange de gaussiennes donnant le maximum de vraisemblance (ce qui revient à approcher la probabilité *a posteriori* maximale, vu qu'on a considéré les probabilités *a priori* des gaussiennes comme égales).

Une fois les GMM construits à partir de données d'entraînement de chaque classe, le classement des sujets tests se fait par le calcul de la vraisemblance de leurs données par rapport aux différents modèles GMM. La classe attribuée est généralement celle du modèle GMM qui correspond à la plus grande vraisemblance.

3.1.2.2 Modèles discriminants

K plus proches voisins La méthode des k plus proches voisins (ou k-NN) est une des méthodes de classification discriminante les plus simples. La phase d'apprentissage consiste seulement dans le stockage des vecteurs caractéristiques des données d'entraînement et des labels des classes associées. La phase de test consiste à classer les vecteurs non labellisés de données test en fonction du label majoritaire chez les k échantillons les plus proches. L'hyper paramètre k est

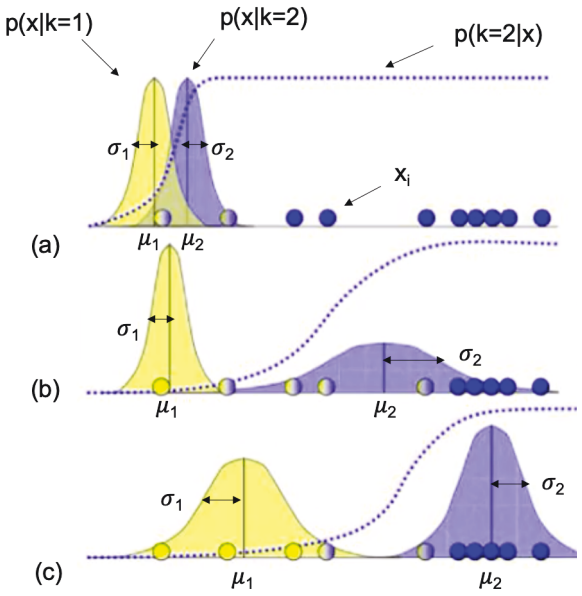


FIGURE 3.4 – Algorithme Espérance-Maximisation (EM). Représentation en 1 dimension avec un mélange de 2 gaussiennes. (a) correspond à l'initialisation : les paramètres μ et σ des 2 fonctions gaussiennes d'indice k=1 et 2, sont initialisés aléatoirement. La probabilité *a posteriori* $p(k|x_i)$ est calculée pour chaque observation x_i (étape d'estimation). Les moyennes et variances des gaussiennes 1 et 2 sont modifiées (b) pour correspondre aux moyennes et variances des observations pondérées par les probabilités *a posteriori* calculées précédemment (étape de maximisation). Les étapes d'estimation et de maximisation sont réitérées jusqu'à convergence (c).

fixé par l'utilisateur et peut être modifié de manière à augmenter le taux de données test bien classées. Cette méthode de classification donne directement le résultat de classification (binaire) sans passer une distribution de probabilité.

Régression logistique La régression logistique est une méthode de classification linéaire discriminante, basée sur une régression binomiale. La régression logistique repose sur l'hypothèse suivante :

$$ln\frac{p(Y=1|X)}{1-p(Y=1|X)} = b_0 + b_1x_1 + ... + b_nx_x \tag{3.7}$$

avec $X = (x_1, ...x_n)$ le vecteur caractéristique (les variables prédictives), Y la classe de valeur 0 ou 1 (la variable à prédire), et b_i les paramètres du modèle à estimer. Cette hypothèse est valable pour plusieurs types de distributions de X, notamment les distributions multinormales et pour le cas où x_i sont des valeurs booléennes (0 ou 1).

La phase d'entraînement consiste à estimer au mieux les paramètres b_i du modèle de régression, à partir de l'estimation du maximum de vraisemblance. La phase de test consiste à calculer les scores de classification $b_0 + b_1x_1 + ... + b_nx_x$ à partir de nouvelles données. Si ce score est positif on considère que Y=1 et sinon que Y=0. On considère cette analyse discriminante par extension, ce ne sont pas les densités de probabilité conditionnelles $p(1|X)$ et $p(0|X)$ qui sont estimées mais leur rapport.

SVM Les machines à vecteur de support, ou séparateur à vaste marge (SVM) sont un autre ensemble de classification discriminante, reposant sur deux idées clefs : la marge maximale et la fonction noyau.

Cas de données linéairement séparables Considérons dans un premier temps un problème de discrimination linéairement séparable, c'est à dire qu'il existe un hyperplan permettant de séparer totalement les données d'entraînement. Autrement dit il existe une fonction de décision linéaire de la forme :

$$D(X) = signe(g(X)) \qquad \text{avec } g(X) = W \cdot X + w_0 \tag{3.8}$$

classant correctement toutes les observations de l'ensemble d'apprentissage. On a alors pour tout les éléments k de l'entraînement :

$$l_k g(X_k) > 0 \qquad \text{avec } l_k \text{ leur label (1 ou -1)} \tag{3.9}$$

Quand les observations sont linéairement séparables, il existe en général une infinité d'hyperplans séparateurs (décrits par $g(X) = 0$). L'enjeu étant de choisir le meilleur séparateur. Soit H un hyperplan séparateur, de paramètres W et w_0. Définissons les vecteurs supports les éléments de l'apprentissage les plus proches de l'hyperplan H. Notons X^+ le vecteur support correspondant à la classe 1 et X^- celui correspondant à la classe -1. Ajustons w_0 de manière à avoir la distance entre H et X^+ égale à la distance entre H et X^-. On définit alors la marge comme étant cette distance. On considère l'hyperplan optimal comme étant l'hyperplan séparateur avec la plus grande marge possible. Un exemple d'observations linéairement séparables en dimension 2 est représenté Figure 3.5 avec la représentation d'un mauvais séparateur linéaire, d'un bon séparateur linéaire et du séparateur linéaire optimal. En dimension 2 les hyperplans séparateurs H (décrits par paramètres W et w_0) sont des droites de vecteur normal $\vec{n} = \frac{\vec{w}}{||\vec{w}||}$ et d'ordonnée à l'origine $d = \frac{w_0}{||\vec{w}||}$.

$$\begin{aligned} Marge &= \min_k(Dist(X_k, H) && (3.10)\\ &= \min_k(\frac{|g(X_k|}{||W||}) && (3.11)\\ &= \frac{1}{||W||}\min_k(l_k g(X_k) && (3.12) \end{aligned}$$

Ajustons $||W||$ de manière à avoir $g(X^+) = 1$ (et par conséquent $g(X^-) = -1$). La marge vaut alors $\frac{1}{||w||}$ et la condition de séparation linéaire décrite par la formule 3.9 devient :

$$l_k g(X_k) \geq 1 \tag{3.13}$$

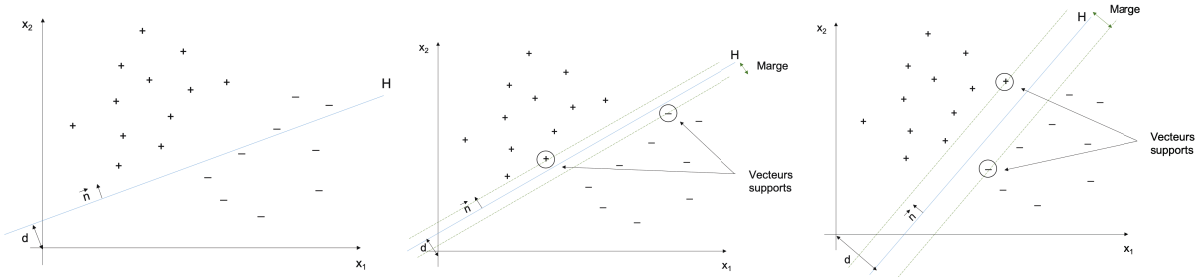
Rechercher l'hyperplan optimal revient donc à chercher le couple de paramètres W et w_0 maximisant $\frac{1}{||W||}$ (ce qui revient à minimiser $\frac{1}{2}||W||^2$) tout en respectant la condition 3.13. On cherche donc à résoudre le système suivant (formulation primale des SVM) :

$$\begin{cases} \min\limits_{W,w_0} \frac{1}{2}||W||^2 \\ avec \quad l_k(W \cdot X_k + w_0) \geq 1 \quad \forall k \in entrainement \end{cases} \tag{3.14}$$

Ce système peut se résoudre par la méthode des multiplicateurs de Lagrange.

Cas non linéaire Dans le cas général, les données ne sont pas linéairement séparables. Afin de remédier à ce problème, on peut considérer ces données dans un espace de dimension supérieure (dit espace de redescription) dans lequel il est probable qu'il existe une séparation linéaire. Pour cela on considère une transformation non linéaire ϕ de X vers un espace de dimension supérieure muni d'un produit scalaire. On va définir une fonction noyau k à valeur dans $\Re$ à partir de ce produit scalaire :

$$k(X, X') = \phi(X) \cdot \phi(X') \tag{3.15}$$



(a) Mauvais séparateur linéaire (b) Bon séparateur linéaire (c) Séparateur linéaire optimal

FIGURE 3.5 – Ensemble de points en dimension 2 linéairement séparables avec 3 exemples d'hyperplans H séparateurs caractérisés par leur vecteur normal $\vec{n}$ et leur ordonnée à l'origine d) . En a) H est un mauvais séparateur linéaire car certains points sont mal classés par ce séparateur (l'équation 3.9 n'est pas vérifiée). En b) H est un bon séparateur linéaire (l'équation 3.9 est vérifiée). En c) H est le séparateur linéaire optimal (l'équation 3.9 est vérifiée et la marge est maximale)

.

En pratique on ne connaît pas ϕ, on construit directement la fonction noyau, à partir des conditions que doit remplir un produit scalaire. La fonction noyau k doit être une fonction symétrique, semi-définie positive. Elle associe à tout couple d'observations (X,X') de ($\Re^n$,$\Re^n$) une mesure de leur « influence réciproque » calculée généralement à travers leur produit scalaire dans $\Re^n$ (noyaux projectifs) ou leur distance (noyaux radiaux). Dans le cas du noyau linéaire $k(x, x') = X \cdot X'$ on se ramène au cas du classifieur linéaire sans changement d'espace.

Des exemples typiques de noyaux utilisés dans le cas non linéaire sont :

- le noyau polynomial : $k(x, x') = (\sigma + X \cdot X')^p$ (3.16)

- le noyau gaussien : $k(x, x') = e^{\frac{||X - X'||^2}{2\sigma^2}}$ (3.17)

La plupart des noyaux dépendent d'un paramètre σ, appelé largeur de bande, dont le réglage est souvent critique pour le bon fonctionnement de la méthode.

Au final la fonction décrivant le séparateur est une combinaison linéaire de noyaux dont le signe final correspond à la classe.

Cas des données non séparables En général les données ne sont pas parfaitement séparables, que ce soit avec un séparateur linéaire ou avec la méthode du noyau. Un classement est néanmoins possible avec la technique dite de marge souple qui tolère les mauvais classements. Cette technique consiste à trouver le meilleur compromis entre le nombre d'erreurs de classement et la largeur de la marge. On définit ξ_k la variable d'écart liée à l'observation X_k telle que :

$$\xi_k = max(0, 1 - l_k(W \cdot X_k + w_0)) \tag{3.18}$$

Si l'observation X_k est bien classée par le séparateur alors la condition 3.13 est vérifiée et donc $\xi_k = 0$. Si X_k est du mauvais côté du séparateur alors $\xi_k = 1 - l_k(W \cdot X_k + w_0) > 0$ représente son erreur. On cherche à maximiser la marge tout en minimisant la somme des erreurs pondérée par un terme C positif d'équilibrage choisie par l'utilisateur. La formulation primale SVM devient alors :

$$\begin{cases} \underset{W,w_0,\xi_k}{min} \frac{1}{2}||W||^2 + C\sum_{k=1}^{n}\xi_k \\ avec \quad l_k(W \cdot X_k + w_0) \geq 1 - \xi_k, \;\; \xi_k > 0, \quad \forall k \in entrainement \end{cases} \tag{3.19}$$

Réseaux de neurones Un autre grand champ de modèles discriminant est celui des réseaux de neurones artificiels. La construction de ces classifieurs est inspirée du fonctionnement des neurones biologiques. Ces réseaux se déclinent sous de multiples formes et sont très utilisés dans le domaine de reconnaissance de formes. Par soucis de concision nous ne détaillerons par leurs principes de fonctionnement dans cette section.

3.1.3 Evaluation de la performance

Plusieurs métriques peuvent être utilisées pour évaluer la performance d'un modèle. Si on fixe le seuil de classification *a priori*, le plus courant est d'utiliser pour évaluer la performance l'accuracy(Acc), qui est définie comme étant le taux de bonnes classification (nombre de sujets bien classés sur nombre de sujets testés). Cette mesure est souvent accompagnée de la Sensibilité (Se), traduisant le taux de bonnes classifications parmi les sujets malades, et de la Spécificité (Sp), correspondant au taux de bonnes classifications parmi les sujets sains.

$$Acc = \frac{nombre\ de\ sujets\ bien\ classes}{nombre\ de\ sujets\ testes}$$

$$Se = \frac{nombre\ de\ sujets\ MP\ bien\ classes}{nombre\ de\ MP\ testes}$$

$$Sp = \frac{nombre\ de\ sujets\ sains\ bien\ classes}{nombre\ de\ sains\ testes}$$

On peut parler également de faux positifs (nombre de sujets sains classés comme MP) et de faux négatifs (nombre de MP classés comme sains). Une sensibilité grande signifie donc peu de faux négatifs et une spécificité élevée peu de faux positifs. A l'inverse les termes de vrais positifs et de vrais négatifs traduisent respectivement le nombre de MP bien classés et le nombre de sains bien classés. La matrice de confusion fait la synthèse des valeurs, ou des taux, de vrais et faux négatifs et de vrais et faux positifs, cf. Tableau 3.1.

	classés MP	**classés sains**
MP réels	vrais positifs	faux négatifs
sains réels	faux positifs	vrais négatifs

TABLE 3.1 – Matrice de confusion

On peut également visualiser des vrais et faux négatifs et vrais et faux positifs en traçant la distribution des scores des sujets avec le seuil choisi pour la classification, cf. Figure 3.6

On peut également choisir de modifier le seuil *a posteriori*, afin d'obtenir l'équilibre que l'on souhaite en spécificité et sensibilité. Un seuil couramment retenu est celui pour lequel $Se = Sp$, le taux d'erreur correspondant est appelé Equal Error Rate (EER). On peut également choisir le seuil qui minimise le nombre total d'erreurs de classification, le taux de bonnes classifications obtenu est appelé maximum accuracy.

Si l'on souhaite avoir une idée plus complète des performances d'un modèle on peut calculer le taux d'erreur en faisant varier le seuil (du min au max). La courbe ROC correspond au tracé de Se en fonction de $1 - Sp$, c'est à dire au taux de MP bien classés (taux vrais positifs) en

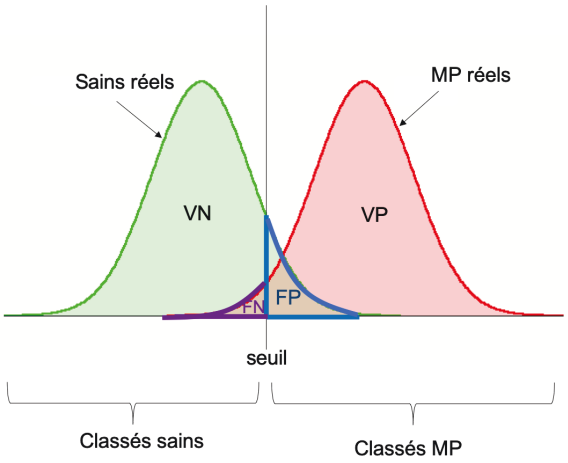
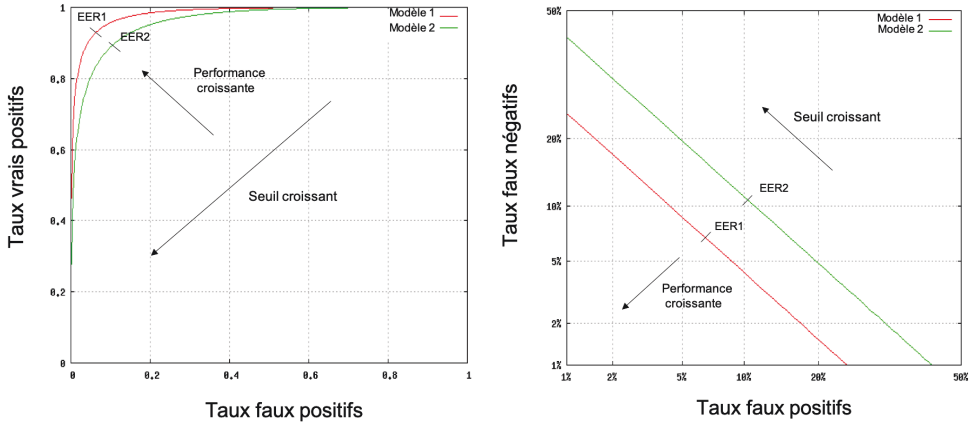


FIGURE 3.6 – Distribution des scores de sujets MP et sains et seuil choisi pour la classification. FP : faux positifs, FN : faux négatifs, VP : vrais positifs, VN : vrais négatifs

.

fonction du taux de sains mal classés (taux faux positifs). L'air sous la courbe (AUC pour *Area Under the Curve*) traduit la performance globale du système de classification (indépendamment du seuil). Plus l'AUC est important plus le système est performant. On peut également tracer le taux de faux négatifs ($1 - Se$) en fonction du taux de faux positifs ($1 - Sp$). Si on utilise une échelle non linéaire dite *normal deviate mapping*, la courbe obtenue est plus linéaire que les courbes ROC, et met plus en valeur les différences autour du point (seuil) de fonctionnement. Cette courbe est appelée courbe DET pour *Detection Error Tradeoff*.



(a) Courbes ROC (b) Courbes DET

FIGURE 3.7 – Courbes ROC à gauche et DET à droite associées à deux modèles de classification, le modèle 1 (rouge) étant plus performant que le modèle 2 (vert).

.

3.1.4 Validation

Quand beaucoup de données sont disponibles, la meilleure approche est de séparer la base de données en 3. Une première partie (données d'entraînement) sera dédiée à l'entraînement des modèles, en trouvant les meilleurs valeurs pour les paramètres des modèles (moyennes et SD des gaussiennes pour les GMM, poids des neurones pour les réseaux de neurones..). Une deuxième partie de la base, appelée base de test, sera utilisée pour tester ces modèles et évaluer

leurs performances. Ces deux parties de la base sont souvent considérées ensemble en tant que base de développement. Les hyperparamètres (nombre de gaussiennes pour les GMM, nombre de couches pour les réseaux de neurones...) pourront être ajustés, grâce aux différents tests effectués, de manière à optimiser la performance. La performance obtenue pour chaque modèle permettra de choisir le meilleur modèle. L'erreur de test est généralement une estimation optimiste de l'erreur réelle (ou de généralisation) car elle peut contenir du surapprentissage. C'est pourquoi il est préférable d'utiliser la troisième partie de la base de données, dite base de validation, pour estimer l'erreur réelle du modèle choisi, c'est à dire l'erreur de prédiction qu'on aurait sur une infinité de nouvelles données. La nomenclature peut varier d'un article à l'autre, certains appellent base de test ce que nous appelons base de validation et réciproquement.

Le nombre de sujets mal classés peut être considéré comme la somme de variables aléatoires de Bernouilli de paramètre p, avec p la probabilité réelle de mal classer un sujet. Si on considère que les n sujets testés de la base de validation sont indépendants et identiquement distribués (iid), alors le nombre de sujets de validation mal classés suit une loi binomiale d'espérance np et de variance $np(1-p)$.

Le taux d'erreur de validation Err_{val} (nombre de sujets mal classés sur nombre de sujets testés) suit alors également une loi binomiale, dont la moyenne est p et la variance $p(1-p)/n$.

Si $np(1-p) > 5$, c'est à dire n suffisamment grand et p suffisamment éloigné de 0 et 1, alors cette loi binomiale peut être estimée par la loi normale de même moyenne et même variance.

De même on peut aussi estimer le taux d'erreur réel (p) par une loi normale de moyenne égale à Err_{val} et de variance définie par :

$$variance = Err_{val}(1 - Err_{val})/n \tag{3.20}$$

C'est à dire qu'on représente la densité de probabilité du taux d'erreur réel par une fonction gaussienne centrée autour de l'erreur de validation Err_{val}, et de variance définie en 3.20. On peut remplacer dans les formules précédentes Err par Acc, Se (n est alors remplacé par le nombre de sujets MP testés), Sp (n est alors remplacé par le nombre de sujets sains testé) ou EER (n est remplacé par min[nombre MP, nombre sain]). L'intervalle de confiance à 95% de l'estimation de l'erreur réelle est défini par :

$$Err_{val} \pm 1.96\sqrt{Err_{val}(1 - Err_{val})/n} \tag{3.21}$$

Ce qui signifie que l'erreur réelle a une probabilité de 95% d'être dans cet intervalle.

Si on effectue la validation sur un ensemble infini de données de validation indépendantes de taille n, alors on appelle biais l'écart entre la moyenne des taux d'erreurs de validation et le taux d'erreur réel. Quant à la variance associée à l'estimation du taux d'erreur réel, telle que définie dans l'équation 3.20, elle correspond à la variance des taux d'erreurs de validation. Si les bases de validation sont composées d'éléments indépendants et identiquement distribués (c'est à dire représentant bien la population générale) alors normalement le biais est nul. Concernant la variance, plus la taille des bases de validation est grande, plus elle est faible.

Dans le cas où on n'a pas suffisamment de données pour les séparer en trois grandes bases, des techniques de validation croisée ou de *bootstrap* permettent d'estimer l'erreur réelle en minimisant le risque de surapprentissage. Le principe repose sur des méthodes d'échantillonnage. L'idée est de partitionner la base de données en données d'entraînement et données de test, afin d'entraîner le modèle sur la base d'entraînement et de le tester sur la base de test. Cet échantillonnage est répété plusieurs fois (on parle de *runs*). L'erreur réelle est définie ici comme l'erreur qu'on aurait si on entrainait le modèle à partir de toute notre base de données (soit notre base de développement) et qu'on le testait sur une infinité de nouveaux sujets. Son estimation

est basée sur les résultats de classification obtenus pour chaque run.

Le **Leave One Subject Out** (LOSO) est une méthode de validation croisée consistant à prendre pour chaque run un seul sujet pour le test et le reste des sujets pour l'entraînement. Le nombre de runs correspond au nombre de sujets, de manière à ce que tous les sujets aient été testés une et une seule fois. La décision de classification des sujets se fait à l'issue du run où ils sont testés. L'erreur de validation croisée Err_{CV} est le taux d'erreur final une fois que tous les sujets ont été testés. Bien que les tests ne soient pas complètement iid (le sujet testé dans un run intervient dans l'entraînement des autres runs), [Kohavi, 1995] montre que si le classifieur est stable alors l'estimation de l'erreur réelle par Err_{CV} est non biaisée, sa variance peut être calculée par la formule 3.20 et son intervalle de confiance à 95% par la formule 3.21. On appelle classifieur stable un classifieur pour lequel les résultats de classification ne changent pas quand les données sont légèrement modifiées.

La méthode LOSO fait partie des méthodes de validation croisée de type **k-fold**. Les méthodes k-folds consistent à séparer la base en k groupes (ou plis) disjoints, chaque pli étant testé lors d'un run différent. LOSO est le cas particulier où k correspondant au nombre de sujets dans la base. Tout comme le LOSO les décisions de classification se font à l'issu de chaque run, Err_{CV} étant le taux d'erreur final une fois tous les sujets testés. La décision de classification dépend du seuil choisi pour chaque run. On peut choisir un seuil *a priori* (0 par exemple si le score est un LLH ratio, 0.5 si ce score est normalisé), ou choisir le seuil *a posteriori*, correspondant à l'EER par exemple. La modification du seuil *a posteriori* n'est pertinente que s'il y a suffisamment de sujets tests de chaque classe dans le run, donc ne peut pas par principe s'appliquer au LOSO. Choisir le seuil *a posteriori* nécessiterait dans l'idéal une validation ultérieure sur une nouvelle base avec le seuil choisi précédemment. Les mêmes formules que pour LOSO pour estimer l'erreur réelle et sa variance à partir de Err_{CV} peuvent s'appliquer. Néanmoins il faut noter que si k est trop faible alors la perturbation induite par la diminution importante des données d'entraînement risque d'entraîner un biais pessimiste (l'erreur estimée risque d'être plus grande que l'erreur réelle) [Kohavi, 1995]. La variance quant à elle ne dépend pas de k. Habituellement les études choisissent k=10, ce qui a l'avantage d'être moins coûteux en calcul que le LOSO sans trop biaiser l'estimation de l'erreur.

Une autre méthode de validation croisée consiste à partitionner aléatoirement les sujets sans remise en base d'entraînement et base de test, et de réitérer ce partitionnement un certain nombre de fois. Cette méthode est appelée **repeated random subsampling**, ou méthode de Monte Carlo. Les sujets ne sont pas testés le même nombre de fois, et il y a un risque que certains sujets ne soient pas testés s'il n'y a pas assez de runs. Err_{CV} est définie ici comme étant la moyenne des Err de chaque run et la variance de l'erreur réelle est estimée par rapport à la variance des Err de chaque run [Kohavi, 1995]. On aurait divisé cette variance par le nombre de runs si les bases de tests de chaque run avaient été iid entre elles, ce qui n'est pas le cas car elles peuvent avoir des sujets en commun. L'intervalle de confiance à 95% de l'estimation de l'erreur réelle peut être estimé à $Err_{CV} \pm 1.96 * SD$ avec SD l'écart type de l'erreur des runs.

Pour avoir une idée plus générale de la performance de cette validation croisée, on peut moyenner d'autres mesures de performances calculées sur chaque run, comme la sensibilité ou la spécificité en gardant le même seuil de classification pour tous les runs [Sáenz-Lechón et al., 2006]. On peut également moyenner les courbes DET, ROC et moyenner l'AUC. Pour les mesures qui dépendent directement d'un seuil en particulier, comme l'EER ou la maximum accuracy, on peut soit les calculer à partir des courbes DET moyennées [Sáenz-Lechón et al., 2006] (donc en ajustant le seuil sur la DET moyennée), soit les calculer pour chaque run (donc ajuster le seuil pour chaque run) puis les moyenner [Gómez-Vilda et al., 2017].

La méthode de validation croisée **repeated random subsampling** a l'avantage qu'on peut choisir le nombre de runs indépendamment des proportions choisies pour le partitionnement. Elle permet également de pouvoir choisir des proportions différentes suivant les classes, ce qui permet par exemple d'avoir un même nombre de sujets MP et sains pour l'entraînement de chaque run, même si le nombre de sujets MP et sains diffère dans la base totale.

Un échantillonnage aléatoire, pour chaque run, mais cette fois-ci avec remise est aussi une méthode possible de validation, ce sont les méthodes **bootstrap**. Les bases d'entraînement font la même taille que la base totale, mais plusieurs sujets y figurent en plusieurs exemplaire, on estime qu'elle contient en moyenne deux tiers des sujets de la base totale. L'erreur *out-of-bootstrap* est calculée sur les sujets ne figurant pas dans la base d'entraînement. La moyenne pour chaque run donne une estimation fiable de l'erreur réelle, et la variance est calculée de la même manière que pour le *repeated random subsampling*.

Dans la majorité des études, la validation croisée est effectuée plusieurs fois en faisant varier les hyperparamètres afin d'obtenir le meilleur Err_{CV} possible. Même si la validation croisée réduit le risque d'overfitting, avec cette méthode elle ne le supprime pas, surtout si le nombre d'hyperparamètres qu'on fait varier est important. Afin d'avoir une estimation précise et non biaisée de façon optimiste de l'erreur réelle, il faudrait soit tester à la toute fin le modèle choisi sur une nouvelle base de données, soit procéder à une **validation croisée imbriquée**. Cette dernière méthode consiste à partitionner pour chaque run la base en 3. Une partie entraînement servant à la construction du modèle, une partie test servant à l'optimisation des hyperparamètres et une dernière partie servant à l'évaluation de la performance, une fois les meilleurs hyperparamètres choisis. Err_{CV} donnerait alors une estimation non biaisée de l'erreur réelle, qui serait obtenue en prenant la moyenne des meilleurs hyperparamètres de chaque run. Dans la pratique le nombre restreint de sujets disponibles dans les bases de données pour la classification MP vs sain rend difficile le fait de garder des sujets pour la validation finale, ainsi que la validation croisée imbriquée. En effet l'optimisation des hyperparamètres se fait à partir des résultats des sujets testés par run, qui constituent souvent un ensemble trop réduit pour avoir une bonne évaluation de la performance résultant du run en question. Néanmoins la validation croisée simple permet tout à fait de faire de la sélection de modèles, des précautions d'interprétations sont seulement à prendre pour l'estimation de l'erreur réelle du modèle final.

3.1.5 Méthodes ensemblistes

Les méthodes ensemblistes utilisent conjointement plusieurs algorithmes d'apprentissages pour améliorer les performance prédictives que ce soit pour une régression ou pour une classification. Plusieurs types de méthodes ensemblistes existent, nous détaillerons le **b**ootstrap **agg**regat**ing** (ou bagging), introduite en 1196 par Breiman [Breiman, 1996] et ses variantes.

Le **bagging** s'appuit sur l'échantillonnage bootstrap. Il consiste à moyenner la prédiction sur un ensemble d'échantillons bootstrap, réduisant ainsi sa variance. Si on prend un sujet extérieur à notre base de données et qu'on veut le classer avec une méthode bagging, il suffit d'opérer la classification à chaque run en utilisant le modèle entrainé avec l'échantillon bootstrap correspondant au run. La décision finale pourra correspondre au vote majoritaire par exemple (la classe la plus souvent attribuée au cours des runs). Si le score de classification (la probabilité de classe) est disponible pour chaque run, la classification finale peut aussi se faire à partir de la moyenne de ces scores. Ceci améliore l'estimation de la probabilité de classes et produit des classifieurs agrégé avec une variance plus faible [Friedman et al., 2001]. Les modèles agrégés (modèles formés à partir de l'agrégation des modèles de chaque run) ont généralement une meilleure performance de classification que le modèle simple, surtout si le classifieur est instable.

Les modèles agrégés sont souvent utilisés avec comme classifieur des arbres décisionnels, mais ils ont également été utilisés en reconnaissance du locuteur à partir de GMM [Andrews et al., 2000] et de quantification vectorielle [Kyung and Lee, 1999], où ils ont montré de meilleures performances de classification que le modèle simple correspondant.

L'erreur *out-of-bag*, calculée en moyennant les prédictions des sujets sur les échantillons bootstrap où ils ne sont pas sélectionnés, est une bonne estimation, non biaisée, de l'erreur réelle du classifieur agrégé, de la même façon que l'Err_{CV} du LOSO estime l'erreur réelle du classifieur simple entrainé sur la totalité de la base [Friedman et al., 2001, Vaiciukynas et al., 2017]. Au même titre que le LOSO on peut utiliser la formule 3.20 pour calculer la variance associée à l'estimation de l'erreur réelle.

Des variantes du *bagging*, reposant sur de l'échantillonnage sans remise (comme le *repeated random subsampling*) ont également été developpées [Bühlmann and Yu, 2002, Maillard et al., 2017] et montrent les mêmes performances que le bagging. L'erreur équivalente à l'erreur *out-of-bag* est alors calculée en moyennant les prédictions des sujets à chaque fois qu'ils ont été dans le groupe test.

Une autre variante connue consiste à pondérer la moyenne des prédictions par des poids associés à chaque run en fonction de sa performance de classification, cette méthode s'appelle le **boosting**.

Tout comme pour la validation croisée, optimiser les hyperparamètres en fonction de l'erreur *out-of-bag* peut entraîner de l'overfitting. Pour la partie sélection et comparaison de modèles, cela ne constitue pas un problème, mais pour avoir une estimation vraiment fiable de l'erreur de généralisation, il faudrait l'estimer à la fin à partir d'une nouvelle base, ou optimiser les hyperparamètres à chaque run, et calculer la prédiction sur quelques sujets qu'on aurait isolés pour ce run.

3.2 Classification MP vs sain à partir des paramètres globaux

La plupart des études concernant la détection de Parkinson par l'analyse de la voix ont utilisé des approches d'analyse globale. Elles se sont intéressées à des paramètres de temps longs : les auteurs ont dénombré certains événements (nombre de pauses, de dysfluences) qui avaient lieu durant une tâche vocale, ou ont moyenné des paramètres locaux, calculés sur des fenêtres temporelles de l'ordre de 50ms (shimmer, jitter, ratio signal sur bruit, formants, durée des consonnes etc) [Little et al., 2009, Rusz et al., 2011a, Sakar et al., 2013, Rusz et al., 2013a, Novotný et al., 2014].

Une fois les paramètres acoustiques vocaux extraits, des tests de significativité sont effectués (comme le test de Student pour les paramètres avec distribution gaussienne, et le test non paramétrique de Wilcoxon-Mann-Whitney pour les distributions non gaussiennes [Rusz et al., 2015]) pour évaluer la différence entre les groupes (patients parkinsoniens vs sujets sains par exemple). Dans la plupart des cas, seuls les paramètres significatifs (souvent définis comme tels pour $p < 0,05$) sont gardés. Ensuite des mesures de corrélations (coefficient de corrélation de Bravais-Pearson pour les distributions gaussiennes et corrélation de Spearman pour les données non normalement distribuées [Rusz et al., 2015]) sont effectuées pour éliminer les paramètres redondants. Dans certaines études une analyse séquentielle de Wald est utilisée pour trouver les paramètres vocaux les plus souvent affectés dans la maladie de Parkinson et pour évaluer l'étendue des perturbations vocales pour chaque patient [Rusz et al., 2015, Rusz et al., 2011a].

Auteurs	Nb sujets	Tâches	Paramètres les plus discriminants	
Harel & al. 2004	4 MP 4 SC	- DDK (papapa) - lecture phrase - monologue	- F0 SD (monologue) - durée pause (lecture)	
Rusz & al. 2011	23 MP 23 SC	- voyelles soutenues /a/, /i/, /u/ - DDK(pataka) - lecture texte - lecture texte avec accentuations - lecture phrases émotionnelles - monologue 90 sec	- F0 SD (monologue et phrases émotionnelles) - RIRV(DDK)	
Rusz & al. 2011	24 MP 22 SC	- voyelle soutenue /a/ - DDK(pataka) - monologue 80 sec	- F0 SD (monologue), SPLD (DDK), RFPC (DDK), NHR ("aa") → Acc : 85% - F0 SD (monologue) seul → Acc : 81,3%	
Rusz & al. 2013	20 MP 15 SC	- voyelles soutenues /a/ /i/ /u/ - lecture de phrases - répétitions d'une même phrase - monologue	- VSA (monologue) → Acc : 80,4% - F2i/F2u (monologue) → Acc : 80,0%	
Novotny & al. 2014	24 MP 22 SC	DDK(pataka)	- VOT (pa) SNR, CST (ka), VSQ30, 2FT (ta), DDK rate → Acc : 88% - VOT seul → Acc : 80%	
Rusz & al. 2015	19 MP 19 SC	- voyelle soutenue /a/ - DDK(pataka) - monologue	- VOT (DDK), F0 SD (monol.), NoP (monol.), HNR ("aa") sont les plus discriminants - F0 SD (monol.), Int SD (monol.), NoP (monol.) → Se : 99,1 % , Sp : 87,5%	
Skodda 2015	50 MPt 32 SC	répétition de /pa/ ou /pa-ti/ à un rythme choisi ou imposé	- COV - pa-ti ratio	
Huh & al. 2015	29 MP 26 MSA 37 SC	- voyelle soutenue /a/ - lecture de phrases	"aa"	- F0 (hom.) (MSA/SC et MSA/MP→ Se : 57, Sp : 87) - MPT (fem.)(MSA/SC et MP/SC)
			lecture (hom.)	- F0 SD et PRww (MSA/SC et MP/SC) - TSR et TPT (MSA/SC et MSA/MP)→ Se :64, Sp :73
Harel & al. 2004	1 pré-MP 1 SC	- monologue *étude longitudinale*	- F0 SD (à partir de 5 ans avant le diagnostic)	
Postuma & al. 2012	78 RBD 39 SC	- monologue (UPDRS item18) *étude longitudinale*	analyse perceptive → différences à partir de 7 ans avant diagnostic pour MP et 15 ans pour DCL	
Rusz & al. 2015	16 RBD 16 SC	- voyelle soutenue /a/ - DDK(pataka) - monologue	- DDK reg, RFA, DUV, PDW → Se : 96 % , Sp : 79% - DDK rate et DUV encore + discriminants quand UPDRS>4	
Rusz & al. 2015	19 MP 16 RBD	- voyelle soutenue /a/ - DDK(pataka) - monologue	MP/RBD : - VOT (DDK), F0 SD (monol.), NoP (monol.)	

TABLE 3.2 – Inventaire des études sur les marqueurs vocaux qui pourraient contribuer à un diagnostic précoce de la maladie de Parkinson. La 1^{ere} partie fait l'inventaire des études sur des parkinsoniens récemment diagnostiqués (< 4 ans après diagnostic). La 2^{e} partie concerne la phase préclinique de la maladie de Parkinson. MP : Sujet avec Maladie de Parkinson sans traitement, MPt : Sujet parkinsonien avec traitement, SC : sujet contrôle, pré-MP : parkinsonien dans la phase préclinique, MSA : patient avec atrophie multisystématisée, RBD : sujet atteint de *Rapid eye movement sleep Behaviour Disorder*, DCL : patient avec Démence à Corps de Lévy, DDK : tâche de diadococinésie, UPDRS : *Unified Parkinson's Disease Rating Scale*, Acc : *accuracy*, Se : sensibilité, Sp : spécificité. Pour les abréviations des paramètres acoustiques, se référer au Tableau 2.2.

Les paramètres ainsi présélectionnés sont ensuite utilisés par des algorithmes de classification qui vont tester les différentes combinaisons possibles afin de trouver la combinaison de paramètres qui permette de classer au mieux, de façon automatique, les sujets (patients parkinsoniens vs sujets sains par exemple). Les algorithmes de classification les plus souvent utilisés dans ces études sont des algorithmes à apprentissage supervisé de type machines à vecteurs de support (SVM) associés à des méthodes de réseaux à fonctions de base radiales [Rusz et al., 2015, Novotný et al., 2014].

Comme on peut le voir dans le Tableau 3.2, les études sur le diagnostic précoce de la maladie de Parkinson par l'analyse acoustique de la voix obtiennent des taux de réussite (Acc) de 80% quand elles ne prennent en compte qu'un paramètre spécialement pertinent, par exemple VOT dans la tâche DDK [Novotný et al., 2014], F0 SD dans le monologue [Rusz et al., 2011b], VSA ou F2i/F2u toujours dans le monologue [Rusz et al., 2013a]. Ces mêmes études améliorent leur taux de réussite jusqu'à 85% [Rusz et al., 2011b] voire 88% [Novotný et al., 2014] quand elles prennent en compte plusieurs paramètres acoustiques. Dans une autre étude [Rusz et al., 2015], les auteurs ont obtenu une sensibilité de 99,1% et une spécificité de 87,5% pour un classement parkinsonien débutant vs sujet sain. Les mêmes auteurs ont proposé une classification RBD vs sujet sain avec une sensibilité de 96% et une spécificité de 79%. Il est aussi possible de séparer les patients MP débutants des MSA débutants, mais les sensibilités obtenues sont un peu moins bonnes (de l'ordre de 60%) [Huh et al., 2015].

Les taux de réussite et les mesures de sensibilité et spécificité publiés sont à prendre avec précaution car les études en question n'ont concerné que des groupes de petite taille (une vingtaine de sujets maximum par groupe). De plus les bases de données ne sont généralement pas publiques, ce qui rend difficile les comparaisons.

D'autres études ont réalisé leurs analyses sur des échelles de temps plus courts, en effectuant la classification directement à l'échelle de trames de 20ms. Ces méthodes d'analyses et leurs adaptations dans le cadre de la détection de MP sont présentées dans la partie suivante.

3.3 Classification à partir d'analyses court-terme

Les différentes méthodes d'analyses réalisées sur des échelles de temps courtes (d'une durée moyenne de 20ms) sont inspirées des méthodologies utilisées en reconnaissance du locuteur. Nous commencerons par détailler les techniques de reconnaissances du locuteur, puis nous établirons un état de l'art des études qui ont utilisé ces techniques dans le cadre de la détection de MP.

3.3.1 Cas de la reconnaissance automatique du locuteur

La reconnaissance automatique du locuteur est la reconnaissance automatique d'une personne par l'analyse acoustique de sa voix. Il existe deux applications principales : la vérification du locuteur, qui consiste à vérifier l'identité proclamée par un locuteur, et l'identification du locuteur, qui consiste à reconnaître un locuteur particulier parmi un ensemble connu de locuteurs possibles.

Chaque système de reconnaissance de locuteur a deux phases : la construction de la signature vocale (création d'un modèle génératif) et la vérification (test). Lors de la phase d'apprentissage, la voix du locuteur est enregistrée et un certain nombre de paramètres sont généralement extraits pour former une référence vocale. Dans la phase de vérification, un échantillon de parole est comparé à une ou plusieurs références vocales créées précédemment. Pour les systèmes d'identification, l'énoncé est comparé à plusieurs références afin de déterminer la meilleure correspondance, tandis que les systèmes de vérification comparent l'énoncé à une seule référence

vocale.

Les systèmes de reconnaissance du locuteur peuvent nécessiter que le texte prononcé en phase de vérification soit le même que celui prononcé en phase d'entraînement (système dépendant du texte), ou non (système indépendant du texte). Les systèmes dépendant du texte ont généralement de meilleures performances que ceux qui sont indépendant du texte [Campbell, 1997]. Cette différence peut être expliquée par la variabilité due au contenu linguistique et par la variabilité des durées des enregistrements dans le contexte des systèmes indépendant du texte. Néanmoins ces derniers présentent l'avantage de nécessiter moins de restrictions sur les données et reflètent un contexte d'évaluation plus général.

Nous allons maintenant présenter les techniques les plus utilisées en reconnaissance du locuteur. Nous détaillerons les différentes étapes, allant de la paramétrisation du signal (extraction des paramètres acoustiques, et leur traitement dans le but de diminuer l'effet du bruit et des distorsions) à la classification (présentation des classifieurs les plus utilisés dans ce cadre).

3.3.1.1 Extraction des paramètres cepstraux

Cepstre Comme nous l'avons vu précédemment, la parole, qui est majoritairement composée de sons voisés, peut être décomposée suivant un modèle source-filtre avec la vibration des cordes vocales comme source et le conduit vocal servant de filtre. Le signal sonore $s(t)$ s'écrit alors sous la forme du produit de convolution du signal source $g(t)$ par la réponse impulsionnelle du filtre $h(t)$:

$$s(t) = g(t) * h(t) \tag{3.22}$$

La transformation de Fourier appliquée à ce signal donne l'équation suivante :

$$S(w) = G(w).H(w) \tag{3.23}$$

L'étude seule de la densité spectrale ne permet ainsi pas d'observer la seule contribution du filtre, autrement dit du conduit vocal. Or les informations pertinentes pour la reconnaissance du locuteur, sont justement issues du conduit vocal. C'est pourquoi une transformation supplémentaire du signal est effectuée afin d'isoler les caractéristiques du conduit vocal, il s'agit du cepstre. Le calcul du cepstre de $s(t)$ consiste à prendre la transformée de Fourier inverse du log du module de sa Transformée de Fourier. Il s'écrit sous la forme :

$$c(\tau) = FFT^{-1} log|S(w)| \tag{3.24}$$

avec τ une variable homogène au temps appelée quéfrence.

L'intérêt du cepstre est qu'il permet de décorréler la source du filtre en transformant le produit de convolution en somme :

$$c(\tau) = FFT^{-1} log|G(w)| + FFT^{-1} log|H(w)| \tag{3.25}$$

Le premier terme est alors caractéristique de la source et représente la structure fine du signal fréquentiel S(w). Il apparaît dans le domaine cepstrale par un pic (correspondant au pitch) à haute quéfrence. Le deuxième terme est caractéristique du conduit vocal et représente l'enveloppe spectrale lissée de S(w). Il s'exprime plutôt dans les basses quéfrences. Un simple liftrage (équivalent ceptral du filtrage) passe bas permet de supprimer la contribution de la source. Si on applique de nouveau une TF à ce signal liftré on obtient bien l'enveloppe spectrale lissée du signal sonore, qui correspond à l'information issue du conduit vocal.

Plusieurs types de paramètres peuvent être calculés pour décrire le cepstre, résultant notamment d'une analyse prédictive linéaire (LPC), c'est le cas des paramètres LPCC (*Linear*

Prediction Cepstral Coefficients), et PLP (*Perceptual Linear Prediction*), ou résultant d'une analyse par banc de filtre, comme les MFCC (*Mel Frequency Cepstral Coefficients*).

Nous allons détailler par la suite les paramètres MFCC qui sont les plus répandus dans les systèmes de reconnaissance actuels, et qui sont les paramètres que nous utiliserons pour caractériser le cepstre.

MFCC Les *Mel Frequency Cepstral Coefficients* (MFCC) sont des paramètres acoustiques introduits pour la première fois en 1980 par Davis et Mermelstein [Davis and Mermelstein, 1980] pour la reconnaissance automatique de la parole. Depuis ils sont largement utilisés dans la reconnaissance de la parole et du locuteur. Ces paramètres caractérisent l'enveloppe spectrale de la voix, et reflètent donc la forme et le volume du conduit vocal. Les MFCC exploitent à la fois les propriétés de décorrélation du cepstre et les principes psychoacoustiques de l'oreille humaine. Le calcul de ces coefficients se fait en 5 étapes (préaccentuation, fenêtrage, transformée de Fourier discrète sur chaque trame, filtrage passe bande MEL, log de l'énergie de chaque filtre, transformée en cosinus discrète).

Préaccentuation : L'oreille humaine, à intensité égale, entend mieux les aigus que les graves. Pour refléter cette perception, une première étape de préaccentuation est effectuée (généralement un filtre à réponse impulsionnelle finie du premier ordre), accentuant ainsi les hautes fréquences.

Fenêtrage : Le signal sonore temporel est segmenté en trames de 20 à 40 ms (souvent 25ms pour la reconnaissance de la parole et 20ms pour la reconnaissance du locuteur). On considère le signal vocal comme stationnaire pendant cette durée. Avec une durée plus courte, il y aurait trop peu de signal pour calculer le spectre et sur une durée plus longue, le signal risquerait de ne plus être stationnaire. Un chevauchement entre deux trames consécutives et une multiplication du signal par une fonction de fenêtrage (de type par exemple Hamming, Hann, rectangulaire, Blackman ..) permettent d'éviter les discontinuités entre deux trames consécutives (cf. Figure 3.8).

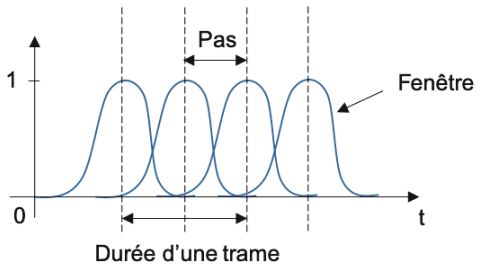


FIGURE 3.8 – Exemple d'allure temporelle de fenêtres.

Transformée de Fourier Discrète (ou DFT pour *Discrete Fourier Transform*) : la DFT est ensuite appliquée pour chaque trame, permettant d'avoir un périodogramme par trame.

Banc de filtres Mel : L'oreille humaine peut être vue comme un ensemble de filtres passe bandes plus espacés pour les fréquences aiguës que graves, impliquant pour l'homme une meilleure discrimination de deux fréquences proches dans les graves que dans les aigus. L'échelle de Mel (dont l'unité est le mel) a été définie pour refléter cette caractéristique. La conversion Mel-Hertz se fait linéairement pour les basses fréquences et logarithmiquement pour les hautes fréquences, selon la formule suivante :

$$Mel(f) = 1127 * ln(1 + f/700) \qquad \text{avec f la fréquence en Hz} \tag{3.26}$$

Dans le but de se rapprocher de l'échelle de la perception auditive, une série de filtres triangulaires dont les largeurs de bandes et l'espacement suivent l'échelle Mel est appliquée aux spectres de chaque trame. L'énergie associée à chaque filtre est calculée (cf. Figure 3.9) créant ainsi un spectre modifié comprenant un nombre réduit de coefficients.

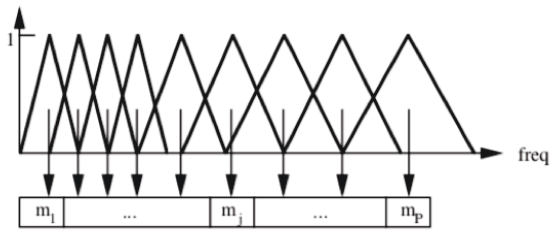


FIGURE 3.9 – Banc de filtres triangulaires Mel. m_i est l'énergie correspondante au ième filtre

Logarithme des énergies : Le logarithme de ces énergies est alors calculé, de manière à suivre la perception humaine logarithmique de l'énergie (intensité) sonore. Le logarithme permet aussi la décorrélation du signal source et du filtre : il transforme le produit de la densité spectrale de la source et du filtre en addition. Il décorrèle aussi les distorsions convolutionnelles linéaires dues à l'environnement d'enregistrement (effet du canal d'enregistrement..), ce qui permet de les supprimer dans un second temps par soustraction moyenne cepstrale.

Transformée en cosinus discrète (DCT pour *Discrete Cosine Transform*) : La DCT est une transformation proche de la transformée de Fourier discrète qui utilise des nombres réels au lieu de nombres complexes. Une DCT inverse est effectuée sur les log énergies issues des filtres MEL. Ce qui a comme effet de regrouper l'information de l'enveloppe spectrale dans ses premiers coefficients (comme le cepstre détaillé dans la partie précédente). De plus la corrélation entre les log énergies consécutives (due entre autres au chevauchement) est supprimée par la DCT. Un des intérêts d'avoir des coefficients décorrélés est de pouvoir utiliser des matrices diagonales de covariances pour les modéliser, via des classifieurs modèle de Markov caché (HMM) ou GMM par exemple, et ainsi réduire la complexité des modèles probabilistes.

$$c_n = \sum_{k=1}^{K} log(m_k) cos(n(k - \frac{1}{2})\frac{\pi}{K}) \tag{3.27}$$

avec n l'indice du coefficient cepstral, k l'indice du filtre MEL, K le nombre de filtres MEL et m_k l'énergie issue du k^{ieme} filtre MEL.

Ces coefficients sont les MFCC. Habituellement entre 12 et 19 coefficients sont gardés pour caractériser l'enveloppe spectrale. Les MFCC extraits forment un vecteur caractéristique par trame. La log énergie totale et/ou le MFCC0 (moyenne des log énergies issues de chaque filtre MEL) peuvent être ajoutés à ce vecteur. Une pondération (via un liftre passe bande) peut également être effectuée. Le poids des premiers et derniers coefficients est alors diminué afin de réduire la sensibilité des coefficients au bruit.

Dérivées temporelles Les dérivées premières (Deltas) et secondes (Delta-Deltas) des MFCC peuvent également être ajoutées au vecteur caractéristique. Les Deltas et Delta-Deltas permettent d'ajouter de l'information quant à l'évolution temporelle des propriétés fréquentielles locales décrites par les MFCC.

3.3.1.2 Traitement du bruit, des distorsions et des pauses

Les performances en reconnaissance du locuteur varient en fonction de la qualité des données voix et de l'environnement dans lequel elles ont été enregistrées. Les vecteurs caractéristiques sont en effet directement affectés par les conditions d'enregistrement. Les techniques de débruitage et de normalisation des paramètres cepstraux visent à réduire les informations spécifiques aux conditions d'enregistrement en affectant le moins possible les caractéristiques spécifiques du locuteur.

Débruitage par Soustraction Spectrale La soustraction spectrale est une méthode de débruitage qui supprime les bruits additifs stationnaires. Soit y(t) un signal sonore bruité composé du signal sonore x(t) non bruité et d'un bruit additif stationnaire b(t). On a :

$$y(t) = x(t) + b(t) \tag{3.28}$$

Ce qui donne sur une trame, après fenêtrage et transformation de Fourier :

$$Y(f) = X(f) + B(f) \tag{3.29}$$

avec Y X et B les transformées de Fourier respectives de y, x et b. Le bruit et le signal non bruité n'étant pas corrélés les densités spectrales d'amplitudes sont liées par l'approximation suivante :

$$|Y(f)| \simeq |X(f)| + |B(f)| \tag{3.30}$$

et les densités spectrales de puissance par :

$$|Y(f)|^2 \simeq |X(f)|^2 + |B(f)|^2 \tag{3.31}$$

Comme le bruit b(t) est stationnaire on peut estimer la densité spectrale de puissance du bruit de fond $|B(f)|^2$ en moyennant celle de trames temporelles correspondant seulement à du bruit de fond : $\overline{|N(f)|^2}$ On peut alors estimer la densité spectrale de puissance du signal non bruité comme :

$$\begin{cases} |X(f)|^2 = |Y(f)|^2 - \overline{|N(f)|^2} & pour\ |Y(f)|^2 - \overline{|N(f)|^2} > 0 \\ |X(f)|^2 = 0 & sinon \end{cases} \tag{3.32}$$

Le calcul de cette densité spectrale non bruitée peut être vue comme une opération de filtrage :

$$|X(f)|^2 = H(f)|Y(f)|^2 \tag{3.33}$$

avec H(f) la réponse fréquentielle du filtre

$$H(f) = \frac{|Y(f)|^2 - \overline{|N(f)|^2}}{|Y(f)|^2} \tag{3.34}$$

$$H(f) = 1 - \frac{1}{SNR + 1} \tag{3.35}$$

avec SNR le rapport signal sur bruit "instantané" (*Signal-to-Noise Ratio*) défini par :

$$SNR(f) = |X(f)|^2 / \overline{|N(f)|^2} \tag{3.36}$$

La soustraction spectrale peut alors être considérée comme une opération de filtrage atténuant la densité spectrale pour les SNR faibles.

Cette opération, quand elle a lieu, est souvent effectuée avant l'étape de filtrage MEL pour le calcul des MFCC.

Normalisation

Soustraction du cepstre moyen Toutes les distorsions du signal ne se limitent pas aux bruits additifs stationnaires. Les canaux d'enregistrement et l'acoustique du lieu peuvent amplifier et diminuer certaines fréquences, agissant comme des filtres stationnaires, ils créent alors des distorsions convolutionnelles stationnaires. Le signal sonore y(t) s'exprime alors comme le produit de convolution de signal vocal x(t) et de la réponse impulsionnelle du canal h(t) :

$$y(t) = x(t) * h(t) \tag{3.37}$$

Après transformation de Fourier et application du log on obtient, pour chaque trame temporelle i :

$$log(Y_i(w)) = log(X_i(w) + log(H_i(w) \tag{3.38}$$

et après application de la transformée de Fourier inverse :

$$Y_i(\tau) = X_i(\tau) + H_i(\tau) \tag{3.39}$$

avec $Y_i(\tau)$, $X_i(\tau)$, $H_i(\tau)$ les cepstres respectifs de $y_i(t)$, $x_i(t)$, $h_i(t)$.
Maintenant soustrayons à $Y_i(\tau)$ la moyenne des cepstres calculés pour chaque trame :

$$Y_i(\tau) - 1/N \sum_{i=1}^{N} Y_i(\tau) = X_i(\tau) - 1/N \sum_{i=1}^{N} X_i(\tau) + H_i(\tau) - 1/N \sum_{i=1}^{N} H_i(\tau) \tag{3.40}$$

Or la distorsion h étant supposée stationnaire, on a :

$$Y_i(\tau) - 1/N \sum_{i=1}^{N} Y_i(\tau) = X_i(\tau) - 1/N \sum_{i=1}^{N} X_i(\tau) \tag{3.41}$$

Si on considère comme négligeables les parties stationnaires de la voix, cette opération de soustraction du cepstre moyen (CMS pour *cepstral mean subtraction*) permet de supprimer la distorsion linéaire convolutionnelle provenant du canal d'information en conservant les informations concernant la voix. Pour cela les portions de parole sur lesquelles sont calculées les moyennes des cepstres doivent être suffisamment grandes et variées en contenu phonétique. La méthode de calcul est alors la suivante :

- calcul des MFCC pour chaque trame ;
- soustraction de la moyenne de chaque MFCC calculée à partir de toutes les trames ;
- division éventuelle par la variance, on parle alors de *Cepstral Mean and Variance Normalization* (CMVN).

Dans le cas où l'effet de distorsion du canal varierait légèrement avec le temps, la CMS peut être effectuée sur des fenêtres glissantes de 3 à 5s. Dans ce cas le vecteur MFCC qu'on normalise se trouve au centre de la fenêtre glissante.

RASTA La méthode RASTA (pour *RelAtive SpecTrAl*) est une généralisation de la CMS, dans le cas de distorsion convolutionnelle, évoluant lentement avec le temps. Cette méthode de normalisation se comporte comme un filtre cepstral supprimant les modulations basses et hautes de fréquences, et non plus seulement la composante continue comme le fait la CMS.

Les méthodes de normalisation par soustraction du cepstre moyen et RASTA ne sont pas nécessaires dans tous les cas, par exemple si un même canal est utilisé. Elles peuvent même détériorer les résultats, par exemple en présence de bruit additif stationnaire, et ne retirent pas les distorsions convolutionnelles non linéaires.

Features warping La technique de normalisation appelée *features warping* est une autre approche de normalisation également utilisée en reconnaissance du locuteur [Pelecanos and Sridharan, 2001]. Elle consiste à appliquer une transformation non linéaire à la distribution des coefficients cepstraux, de manière à la transformer en distribution gaussienne. Cette technique est légèrement plus performante que les précédentes mais est plus coûteuse en calcul.

Features mapping La méthode de *features mapping* est une autre méthode de normalisation, cette fois supervisée, à la différence des méthodes précédentes. Tout d'abord, un GMM indépendant du canal d'enregistrement est créé en utilisant un jeu de données provenant de nombreux canaux différents. Ensuite, des GMM canal-dépendant sont formés en adaptant le GMM canal-indépendant à partir de données provenant d'un même canal. Enfin, les fonctions de mappage sont apprises en examinant comment les paramètres du modèle canal-indépendant changent après l'adaptation. Au cours de l'étape de normalisation, le canal utilisé est d'abord détecté (par exemple par comparaison de la vraisemblance des vecteurs caractéristiques par rapport à chaque GMM canal-dépendant), puis chaque vecteur de caractéristique est mappé dans un espace indépendant du canal à l'aide des fonctions de mappage correspondant au canal. L'intérêt de cette méthode est qu'elle permet de traiter tout type d'effet de canal, et pas seulement les distorsions convolutionnelles linéaires. L'inconvénient est qu'elle nécessite beaucoup de données pour former les différents modèles canal-dépendant et canal-indépendant.

Suppression des pauses : VAD Dans la reconnaissance du locuteur, les pauses ne contiennent pas d'informations intéressantes et peuvent donc être supprimées afin d'améliorer la performance. Cette suppression s'appuie sur la détection de l'activité vocale (VAD pour *Voice Activity Detection*). Généralement on décide qu'une activité vocale est détectée sur une trame si un paramètre particulier (par exemple l'énergie totale ou le MFCC0) dépasse un certain seuil (seuil pouvant dépendre de l'enregistrement entier).

3.3.1.3 Classifieurs utilisés

Modélisation par VQ La quantification vectorielle (VQ), introduite dans les années 1980 [Soong et al., 1985], a été l'une des premières formes de modèles utilisés en reconnaissance automatique du locuteur. Comme détaillé partie 3.1.2.1, son principe réside dans le fait de modéliser la distribution des vecteurs caractéristiques d'une classe par des vecteurs de centroïdes (codebook). La classification des vecteurs tests s'effectue en comparant leur distorsion par rapport aux différents modèles. Pour l'identification du locuteur, un codebook par locuteur est créé à partir des MFCC issus des données d'entraînement. Ensuite, la distorsion entre les données tests d'un sujet et les codebooks est estimée comme étant la moyenne des distances euclidiennes calculées pour chaque trame entre les vecteurs caractéristiques et les codebooks. L'identité du locuteur testé est attribuée à celle qui correspond au codebook pour lequel la distorsion est la plus faible.

Modélisation par GMM En 1992 Reynolds utilise une modélisation probabiliste, plus flexible, basée sur un mélange de gaussiennes (GMM) pour la reconnaissance automatique du locuteur [Reynolds, 1992, Reynolds and Rose, 1995]. Ce modèle détaillé partie 3.1.2.1 permet de décrire plus précisément les distributions des vecteurs caractéristiques.

Pour l'**identification du locuteur**, un modèle GMM est entrainé par locuteur à partir des MFCC issus de leurs données d'entraînement. La vraisemblance (LH pour *likelihood*) des données $X = (x_1, ...x_n)$ d'un locuteur que l'on souhaite identifier par rapport aux K différents modèles de locuteurs est calculée. Si on considère les trames indépendantes, elle vaut alors le

produit des vraisemblances de chaque trame :

$$LH(X, \lambda_k) = P(X|\lambda_k) = \prod_{i=1}^{n} P(x_i|\lambda_k) \tag{3.42}$$

λ_k représentant le modèle du locuteur k et n le nombre de trames.
On considère généralement le log de la vraisemblance (LLH), afin de transformer le produit en somme. Ce log est enfin divisé par le nombre de trames, afin d'avoir un score indépendant de la durée de l'échantillon audio testé. Ce qui revient à calculer la moyenne des log vraisemblances de chaque trame. L'identification attribuée au locuteur testé est alors celle qui correspond au modèle du locuteur pour lequel le score de vraisemblance est le plus élevé.

$$\text{Identification retenue } = \max_{k=1..K} (\sum_{i=1}^{n} \frac{1}{n} logP(x_i, \lambda_k)) = \max_{k=1..K} \frac{1}{n} LLH(X, \lambda_k) \tag{3.43}$$

Pour la **vérification du locuteur**, un GMM est entrainé avec les MFCC issus des données d'entraînement d'un locuteur k et un GMM (appelé modèle du monde) est entrainé à partir de MFCC issus d'un grand ensemble de locuteurs. La log vraisemblance des vecteurs MFCC issus de données du locuteur qu'on veut tester est calculée pour chaque trame par rapport à ces deux modèles, puis moyennée sur l'ensemble des trames. La différence entre ces deux LLH constitue le score S de classification, appelé *log-likelihood ratio*.

$$S(X) = \frac{1}{n} LLH(X, \lambda_k) - \frac{1}{n} LLH(X, \lambda_{monde}) = \frac{1}{n} log(\frac{P(X|\lambda_k)}{P(X|\lambda_{monde})}) \tag{3.44}$$

avec λ_k représentant le modèle du locuteur k et λ_{monde} représentant le modèle du monde.
Si ce score est supérieur à un certain seuil θ, on considère que le locuteur testé est bien celui ayant servi pour l'entraînement du modèle, sinon le locuteur testé est dit "imposteur". Ce seuil de décision est généralement fixé en amont lors d'une phase de développement.

$$\begin{cases} S(X) \geq \theta \text{ alors locuteur testé = locuteur k} \\ S(X) < \theta \text{ alors locuteur testé} \neq \text{locuteur k (imposteur)} \end{cases} \tag{3.45}$$

Des techniques de **normalisation** de scores ont été développées de manière à avoir un seuil indépendant du locuteur. La plupart se fondent sur l'hypothèse que la distribution des scores imposteurs est une distribution gaussienne dont la moyenne μ_{imp} et la variance σ_{imp} dépendent du locuteur k utilisé pour le modèle et/ou du locuteur utilisé pour le test.

$$\Lambda(X) = \frac{S(X) - \mu_{imp}}{\sigma_{imp}} \tag{3.46}$$

μ_{imp} et σ_{imp} peuvent être estimés lors de la phase de développement : un certain nombre d'imposteurs sont testés par rapport au GMM correspondant au locuteur k, résultant en un score moyen μ_{imp} et une variance σ_{imp}. C'est ce qu'on appelle la normalisation ZNorm (*Zero Normalization*).
μ_{imp} et σ_{imp} peuvent aussi être estimés lors de la phase de test : le locuteur que l'on souhaite tester par rapport au modèle k est d'abord testé par rapport à plusieurs modèles correspondant à plusieurs locuteurs (différents de lui). μ_{imp} est alors son score moyen et σ_{imp} sa variance moyenne. C'est ce qu'on appelle la normalisation TNorm (*Test Normalisation*).

Une variante de ZNorm, appelée HNorm (*Handset Normalisaton*) peut être utilisée pour réduire l'effet non linéaire des microphones et canaux dans le cas de conditions d'enregistrements non appariées [Reynolds et al., 2000]. La moyenne et la variance des scores de classification des imposteurs utilisant un même canal et/ou microphone est alors calculée. Le canal et/ou microphone du locuteur test est identifié puis le score du locuteur test est alors normalisé par rapport à la moyenne et variance obtenue précédemment pour ce canal et/ou microphone.

Modélisation par GMM-UBM Les modèles GMM permettent une modélisation simple et robuste de la distribution des vecteurs caractéristiques d'un locuteur, à condition d'avoir assez de données voix pour ce locuteur. Dans le cas où trop peu de données sont disponibles, une méthode, dite GMM-UBM, a été proposée dans [Reynolds et al., 2000]. Au lieu d'entraîner un GMM à partir des seules données voix d'un locuteur, le modèle GMM du locuteur est formé en adaptant un modèle du monde, dit UBM (*Universal Background Model*). L'UBM est un GMM construit à partir de données provenant d'un grand nombre de locuteurs, représentant ainsi une "voix moyenne".

L'adaptation de ce modèle aux données d'un locuteur en particulier est une forme d'adaptation Bayésienne utilisant le critère MAP (Maximum A Posteriori). Tout comme l'algorithme EM (*Expectation Maximization*), cette adaptation se fait en deux étapes. Les paramètres d'initialisation sont les paramètres finaux décrivant le modèle UBM. L'étape d'estimation est identique à celle de EM : les statistiques (qui sont des probabilités *a posteriori*) des vecteurs MFCC du locuteur sont calculées par rapport à chaque gaussienne de l'UBM. L'étape de maximisation est légèrement différente : le calcul des nouvelles moyennes et variances des gaussiennes prend en compte à la fois les données du locuteur avec les statistiques calculées dans l'étape précédente, et les paramètres de l'UBM. Un coefficient d'adaptation, calculé pour chaque gaussienne, régit le compromis entre ces deux dépendances. Pour chaque gaussienne, il privilégie la dépendance liée au locuteur quand beaucoup de données du locuteur sont associées à cette gaussienne, et il donne plus de poids aux paramètres de l'UBM quand peu de données du locuteur y sont associées. Pour construire l'UBM, Reynolds et al. ont utilisé 1h de données voix, à raison de 30s par locuteur, puis l'ont adapté avec 2min de données voix pour chaque locuteur.

La technique GMM-UBM non seulement permet de traiter les cas où la quantité de données voix disponibles d'un locuteur est trop faible pour construire un GMM classique robuste, mais semble, d'après les auteurs, aussi améliorer les performances de classification d'une manière générale de par le couplage des modèles que l'on veut comparer.

Super-vecteurs Plus tard, les méthodes GMM-SVM ont été développées pour la vérification du locuteur [Kharroubi et al., 2001, Campbell et al., 2006]. Le GMM correspondant à chaque locuteur est entrainé par adaptation MAP d'un modèle UBM. Par la suite, les moyennes du GMM résultant sont concaténées pour former un super-vecteur par sujet. Enfin généralement un SVM est entrainé à partir des super vecteurs, de manière à séparer les super-vecteurs correspondant à un locuteur particulier (provenant de différents enregistrements de ce même sujet issus de la phase d'apprentissage) d'une part et ceux provenant d'autres locuteurs d'autre part. La phase de vérification du locuteur consiste alors à classer le super vecteur de ce locuteur, issu de son enregistrement test, avec le SVM entrainé.

Analyses factorielles Des techniques d'analyses factorielles ont été développées par la suite de manière à traiter l'effet du canal sur les super vecteurs et à réduire leur dimensionnalité. Elles reposent sur l'hypothèse que le supervecteur correspondant à une session d'enregistrement (pour un locuteur) peut être décomposé en une composante caractérisant le locuteur et une composante caractérisant le canal d'enregistrement, dont les variabilités résident dans deux sous espaces de dimension réduite. Cette hypothèse reflète l'effet convolutionnel linéaire du canal sur la voix, qui se transforme en effet additif au niveau des MFCC. Plusieurs modèles d'analyses factorielles ont été développés de manière à isoler les caractéristiques du locuteur comme le modèle Eigen Channel [Kenny et al., 2003] le modèle Eigen Voice [Kenny et al., 2005] et le modèle Joint Factor Analysis [Kenny et al., 2007]. En prenant en compte les variabilités dues au canal, ces méthodes ont montré une amélioration des performances dans la reconnaissance du locuteur.

i-vecteurs L'approche i-vecteurs est une extension des modèles d'analyse factorielle proposée en 2011 dans [Dehak et al., 2011] et beaucoup utilisée depuis en reconnaissance automatique du locuteur. A la différence des modèles d'analyse factorielle cités précédemment, elle consiste à isoler un seul sous espace contenant les variabilités à la fois des locuteurs et des sessions. La séparation entre l'effet du locuteur et l'effet de la session ou du canal est alors effectuée dans un second temps, sur l'espace de dimension réduite. Le super vecteur s d'un sujet est décomposé comme suit :

$$s = m + T.w \tag{3.47}$$

Avec m le super vecteur moyen de l'UBM, T la matrice de projection de rang faible, appelée matrice de variabilité totale et w une variable latente de distribution normale, dont le Maximum a Posteriori (MAP) est défini comme étant le vecteur identité ou i-vecteur.

Une fois les i-vecteurs extraits, de manière à compenser l'effet du canal ou de la session, une méthode discriminante, comme l'**analyse discriminante linéaire (LDA)** est souvent effectuée. Elle consiste à projeter les i-vecteurs dans une base orthogonale maximisant la variabilité interclasse (ici interlocuteur) et minimisant la variabilité intraclasse (ici intralocuteur). Cette base, et donc la matrice associée à cette projection, est déterminée suite à un entraînement sur un nombre important de locuteurs avec plusieurs sessions par locuteurs et l'utilisation de différents canaux. La projection peut se faire dans un espace de dimension réduite et la dimension de ce dernier n'excède en principe pas le nombre de classes utilisées pour l'entraînement.

$$\mathbf{y} = A.\mathbf{x} \tag{3.48}$$

Les i-vecteurs $\mathbf{x}$ sont transformés en vecteurs $\mathbf{y}$, après multiplication par la matrice de projection A.

Une étape de normalisation, comme la méthode *Within-Class Covariance Normalisation* (WCCN) complète souvent la LDA. Ensuite la classification, pour la vérification du locuteur, consiste à comparer le i-vecteur (ou i-vecteur projeté après LDA) d'un locuteur test avec le i-vecteur (ou i-vecteur projeté) du locuteur hypothétique. Le locuteur test étant considéré comme le locuteur hypothétique si le score de comparaison dépasse un certain seuil, défini en amont. Pour l'identification du locuteur, le i-vecteur (ou i-vecteur projeté) du locuteur test est comparé à celui de tous les locuteurs hypothétiques et le score maximal indique de quel locuteur il s'agit. La comparaison des i-vecteurs (ou i-vecteurs projetés) peut se faire de plusieurs façons, de manière non probabiliste, comme avec une distance cosinus, ou de manière probabiliste, comme avec la PLDA (*Probabilistic Linear Discriminant Analysis*).

La **distance cosinus** entre deux vecteurs $\mathbf{x}_1$ et $\mathbf{x}_2$ est une mesure de distance simple consistant à calculer le cosinus de l'angle formé entre ces deux vecteurs. Deux vecteurs colinéaires et orientés dans le même sens ont une distance cosinus de 1, et de -1 s'ils sont en sens inverse. La valeur du cosinus s'obtient par le produit scalaire des deux vecteurs divisé par le produit de leurs normes.

$$score = \frac{\mathbf{x}_1.\mathbf{x}_2}{||\mathbf{x}_1|| * ||\mathbf{x}_2||} \tag{3.49}$$

La **PLDA**, introduite en 2007 à l'origine pour la reconnaissance faciale [Prince, 2007], est une version probabiliste de la LDA. Son principe repose sur le fait de décomposer le i-vecteur $\mathbf{x}_{ij}$(du i^{eme} locuteur et de la session j) en un terme moyen $\boldsymbol{\mu}$ calculé sur l'ensemble des locuteurs, un terme $F.\mathbf{h}_{ij}$ dépendant du locuteur, un terme $G.\mathbf{w}_{ij}$ dépendant de la session, et un bruit résiduel $\boldsymbol{\epsilon}_{ij}$ supposé gaussien de matrice diagonale de covariance Σ.

$$\mathbf{x}_{i,j} = \boldsymbol{\mu} + F.\mathbf{h}_{ij} + G.\mathbf{w}_{ij} + \boldsymbol{\epsilon}_{ij} \tag{3.50}$$

Les colonnes de la matrice F représentent la base expliquant la variance interlocuteurs, le vecteur $\mathbf{h}_{ij}$ étant la position du locuteur i dans ce sous espace. Les colonnes de la matrice G représentent la base expliquant la variance intralocuteurs, le vecteur $\mathbf{w}_{ij}$ étant la position du locuteur i pour la session j dans ce sous espace. Durant la phase d'entraînement les paramètres μ, F, G et Σ sont estimés. Durant la phase de test, l'i-vecteur du locuteur testé est comparé à l'i-vecteur du locuteur hypothétique, en estimant la probabilité qu'ils aient la même variable identité h. La PLDA peut être précédée d'une PCA ou d'une LDA afin de réduire en amont la dimensionnalité. L'avantage principal de la PLDA est qu'elle permet des classifications par rapport à des classes (des locuteurs) non utilisées pour l'entraînement.

L'avantage des i-vecteurs par rapport aux méthodes GMM-UBM et GMM-SVM couplées à des modèles d'analyse factorielle comme le *Joint Factor Analysis* (JFA) est l'amélioration des performances en reconnaissance du locuteur, de l'ordre de 4% [Dehak et al., 2011]. L'inconvénient de cette méthode est qu'elle nécessite beaucoup plus de données et de puissance de calcul que les GMM. Afin d'augmenter la quantité de données disponibles pour l'entraînement de la PLDA, [Snyder et al., 2018b] a très récemment montré que dupliquer les données en ajoutant différents types de bruits aux données copiées pouvait améliorer les performances.

L'utilisation des réseaux de neurones pour la reconnaissance de la parole a refait surface depuis une dizaine d'année avec l'arrivée du deep learning. En 2014 [Lei et al., 2014] les ont adaptés pour la reconnaissance du locuteur, en les utilisant à la place des GMM pour la production des i-vecteurs. Ils ont obtenu avec ce système une amélioration relative de 30% par rapport aux i-vecteurs produits avec les GMM.

x-vecteur Dans le même esprit que les i-vecteurs, des études récentes ont utilisé les DNN pour extraire d'autres types d'*embeddings*, c'est à dire d'autres représentations du locuteur. Les auteurs de [Variani et al., 2014] et [Heigold et al., 2016] ont utilisé des d-vecteurs pour la reconnaissance du locuteur dépendante du texte, et plus récemment les auteurs de [Snyder et al., 2016] et [Snyder et al., 2017] ont introduit des x-vecteurs pour la reconnaissance du locuteur indépendante du texte. Les DNN utilisés pour l'extraction des x-vecteurs sont composés de 3 parties représentées Figure 3.10 et détaillées ci dessous :

- Un ensemble de couches *frame-level* prenant comme entrée les vecteurs type MFCC. Ces couches forment un réseau TDNN (Time Delay Neural Network). Les réseaux TDNN sont essentiellement des réseaux *fully connected* qui prennent en compte une fenêtre temporelle glissante.

- Une étape de *pooling* qui concatène les sorties du réseau TDNN sur tout un segment audio (de l'ordre de quelques secondes) en prenant la moyenne et l'écart types des sorties. La sortie de cette étape est une représentation en grande dimension (3000) du segment.

- La dernière partie est un réseau *DNN feed forward* simple prenant comme entrée le vecteur sortant de la couche de *pooling*, réduisant sa dimensionnalité à 512 neurones après la première couche. La dernière couche est une couche de *softmax* donnant les probabilités d'appartenance du segment testé aux locuteurs ayant servi pour l'entraînement du DNN. Les x-vectors sont des vecteurs intermédiaires pouvant être extraits après la première couche (embedding a) ou deuxième couche (embedding b) de cette dernière partie de réseau.

Même si le segment testé n'appartient à aucun locuteur ayant servi à entraîner le DNN, l'extraction du x-vecteur correspondant peut être considéré comme une représentation de ce segment et donc du locuteur. Pour la vérification du locuteur il suffit de comparer ce x-vecteur test avec un x-vecteur résultant d'un enregistrement antérieur du locuteur hypothétique. La

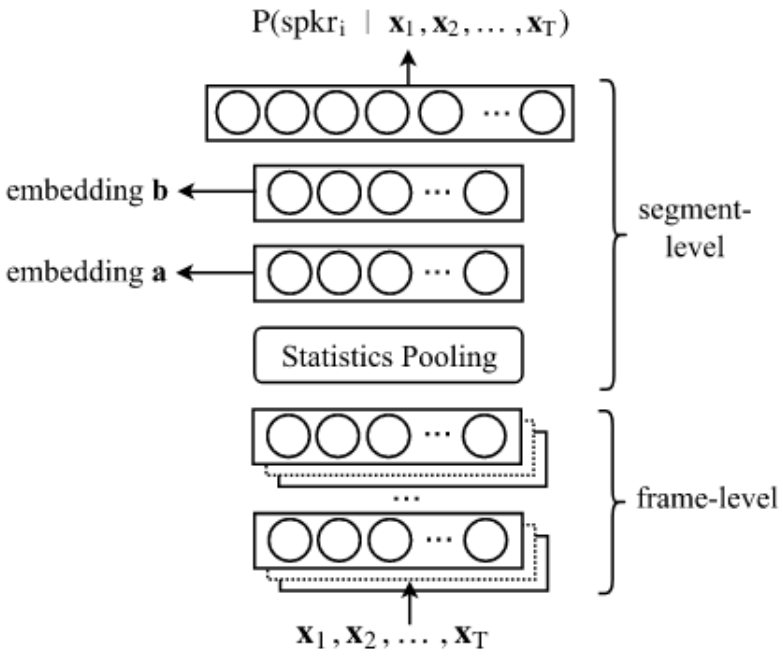


FIGURE 3.10 – Diagramme du DNN proposé dans [Snyder et al., 2017] pour l'extraction d'embeddings appelés x-vecteurs, utilisés pour la reconnaissance du locuteur.

comparaison des x-vecteurs peut se faire, comme pour les i-vecteurs, avec des distances cosinus ou classifieur gaussien (après éventuellement une LDA) ou avec une PLDA.

La méthode des x-vecteurs a amélioré significativement les performances en reconnaissance du locuteur, surtout quand il y a beaucoup de données ou que les données sont augmentées [Snyder et al., 2018b]. De plus les améliorations importantes apportées par les x-vecteurs dans le cas de durées courtes de test et de langues différentes entre l'entraînement du DNN et la partie évaluation [Snyder et al., 2017] semblent indiquer que les x-vecteurs fournissent une représentation plus robuste du locuteur. En 2018 les mêmes auteurs ont montré qu'il était possible d'adapter les x-vecteurs pour la reconnaissance de la langue [Snyder et al., 2018a], en atteignant voire surpassant les performances de l'état de l'art obtenues avec les i-vecteurs.

Réseaux de neurones directs Enfin une dernière méthode consiste à utiliser directement la sortie d'un réseau de neurones pour identifier la classe, avec le score de *softmax*. Cette méthode à été utilisée en identification du locuteur [Parveen and Qadeer, 2000] avec un réseau de neurones récurrents (RNN) et en identification de la langue [Snyder et al., 2018a] avec le DNN servant à extraire les x-vecteurs. La contrainte de ce type de méthode de classification est que la classe hypothétique du fichier test doit forcément faire partie des classes ayant servi à l'entraînement du réseau. Les auteurs de [Snyder et al., 2018a] ont comparé la méthode DNN direct et la méthode x-vecteur pour la détection de la langue et ont trouvé de meilleures performances pour les x-vecteurs.

3.3.2 Détection de MP à partir des MFCC

Les paramètres MFCC sont les paramètres les plus utilisés dans le domaine de la reconnaissance du locuteur depuis des décennies [Bimbot et al., 2004]. Depuis une quinzaine d'années, on commence à les rencontrer dans la détection de pathologies vocales, comme les dysphonies [Dibazar et al., 2002, Godino-Llorente and Gómez-Vilda, 2004, Malyska et al., 2005]. L'utilisation des

MFCC pour la détection de MP a été introduite en 2012 par Tsanas et al. [Tsanas et al., 2012b]. Depuis, plusieurs études ont utilisé les MFCC pour la détection de MP [Godino-Llorente et al., 2006, Bocklet et al., 2013, Benba et al., 2016a, Drissi et al., 2019], en les combinant certaines fois à d'autres paramètres [Orozco-Arroyave et al., 2014b, Orozco-Arroyave et al., 2015a, Hemmerling et al., 2016].

Dans le but d'effectuer la classification MP vs sain, la distribution des MFCC doit être modélisée, à l'échelle du groupe ou à l'échelle du sujet. Plusieurs manières d'effectuer la modélisation sont possibles, dépendant de la forme de la distribution et de la précision souhaitée pour les modèles. Si les MFCC sont groupés autour d'une valeur, impliquant généralement des tâches vocales à faible variété phonétique (comme une voyelle soutenue) et/ou si la précision requise pour le modèle n'est pas trop importante, prendre simplement les moyennes des MFCC, en les combinant éventuellement à d'autres paramètres globaux, peut suffire [Tsanas et al., 2012b, Jafari, 2013, Benba et al., 2014, Benba et al., 2016b].

Si un peu plus d'information est souhaitée pour la description des modèles, plusieurs statistiques peuvent être rajoutées aux moyennes, comme l'écart type, le *kurtosis* (mesure de l'aplatissement) et le *skewness* (mesure de l'asymétrie). Ces modélisations se font toujours à partir de contenu phonétique à faible variété comme les voyelles soutenues [Orozco-Arroyave et al., 2015a, Hemmerling et al., 2016], ou sur des extraits de tâches partageant des caractéristiques acoustiques semblables, comme des trames voisées ou non voisées au sein d'un mot [Orozco-Arroyave et al., 2014b, Orozco-Arroyave et al., 2016a], ou des trames contenant les transitions "non voisé à voisé" [Orozco-Arroyave et al., 2015b].

Pour extraire de l'information de trames qui sont acoustiquement très différentes (comme lors de tâches de lecture ou de monologue), une précision supplémentaire est requise pour décrire les distributions des MFCC. Une modélisation possible est d'utiliser la quantification vectorielle, comme décrit partie 3.1.2.1, en gardant un vecteur codebook par sujet [Benba et al., 2014]. Une autre manière plus précise est de modéliser la distribution des MFCC par des GMM, adaptés d'un UBM, et de garder la moyenne des gaussiennes pour le vecteur caractéristique (méthode GMM-SVM) [Bocklet et al., 2013] . Avec cette méthode de type GMM-SVM (décrite au paragraphe 3.3.1.3) les auteurs ont rapporté un taux de bonnes classifications de 80%.

Au final toutes ces études ont construit un vecteur de caractéristiques par tâche vocale et par sujet (cf. Figure 3.1a). Après avoir généralement effectué des tests de significativité et de redondance pour réduire la dimensionnalité des vecteurs de caractéristiques, ils les ont ensuite introduits dans un classifieur, le plus souvent une machine à vecteur de support (SVM) [Rusz et al., 2011b, Novotný et al., 2014, Benba et al., 2015, Orozco-Arroyave et al., 2015b, Orozco-Arroyave et al., 2016a], mais on aperçoit quelques fois l'usage d'arbres décisionnels [Halawani and Ahmad, 2012, Hemmerling et al., 2016], de *Multilayer perceptrons* (MLP) [Gil and Johnson, 2009, Jafari, 2013], et de la méthode des k plus proches voisins (k-NN) [Ozkan, 2016, Benba et al., 2017, Sakar et al., 2017] . Pour finir ils valident généralement leur classification avec une validation croisée de type k-fold [Tsanas et al., 2012b, Ozkan, 2016] ou repeated random subsampling [Sáenz-Lechón et al., 2006], ou utilisent une méthode ensembliste de type bagging [Halawani and Ahmad, 2012, Vaiciukynas et al., 2017]. Il faut néanmoins prendre certains résultats de performance avec précaution, car dans certaines études [Tsanas et al., 2012b, Jafari, 2013], l'indépendance des locuteurs entre les groupes d'entraînement et de test n'est pas garantie, aboutissant à des performances élevées ($Acc > 95\%$) mais avec de fortes chances d'êtres biaisées de façon optimiste.

Une autre façon de procéder à l'analyse de la voix est de réaliser une analyse dite sur temps court, ou à l'échelle de la trame. Cela consiste à extraire des paramètres calculés sur

des temps courts (de l'ordre de 20 ms) comme les MFCC, et de garder cette fois-ci un vecteur de caractéristiques par trame (cf. Figure 3.1b), la classification se faisant alors directement à l'échelle de la trame. Certains auteurs [Kapoor and Sharma, 2011] ont modélisé la distribution de ces vecteurs de caractéristiques avec une quantification vectorielle de façon à obtenir un modèle Parkinsonien et un modèle sain. Ils ont ensuite calculé la distorsion des MFCC de sujets tests, trame par trame, afin de les classer. Une modélisation plus précise peut être apportée par les GMM, comme détaillé dans notre analyse publiée en 2017 [Jeancolas et al., 2017]. Le calcul de la log-vraissemblance (LLH) par rapport aux modèles GMM MP et sains permettant alors de déterminer la classe des sujets tests à partir de leur vecteurs MFCC. Une méthodologie analogue de modélisation des groupes MP et sains par des GMM-UBM a été utilisée en 2018 par [Moro-Velázquez et al., 2018] et comparée à la technique des i-vecteurs, expliquée partie 3.3.1.3. Pour cette dernière, les auteurs ont effectué une modélisation GMM-UBM des MFCC par sujet, et non par classe, et en ont extrait un i-vecteur par sujet (en utilisant la GPLDA). Ils ont ensuite moyenné les i-vecteurs des sujets MP pour avoir un i-vecteur moyen caractéristique du groupe MP, et ont fait la même chose pour le groupe sain. Une mesure de distance a ensuite été effectuée entre les i-vecteurs des sujets tests et les i-vecteurs des deux classes, afin de déterminer la classe la plus probable. Les performances obtenues avec les i-vecteurs à partir des MFCC se sont révélées similaires à celles obtenues par GMM-UBM (Acc=80%). Les auteurs de [Garcia-Ospina et al., 2018] ont également utilisé des i-vecteurs pour détecter MP. Ils ont extrait un i-vecteur par sujet et ont comparé les méthodes de classification par distance cosinus et de SVM à partir de ces i-vecteurs. Une performance légèrement meilleure a été reportée pour la classification par distance cosinus (Acc=77%).

Récemment, des études se sont aussi intéressées à la détection de MP en utilisant des réseaux de neurones convolutifs à partir des spectrogrammes [Vásquez-Correa et al., 2017a, Zhang et al., 2018, Khojasteh et al., 2018] les performances reportées allant de 75% à 90% de taux de bonne classification.

Certaines études ont utilisé d'autres types de paramètres cepstraux que les MFCC. Les auteurs de [Moro-Velázquez et al., 2018] ont trouvé une amélioration des performances avec les RASTA-PLP, comparé aux MFCC, en utilisant la méthode des i-vecteurs. Les auteurs de [Benba et al., 2017] ont utilisé les HFCC (Human Factor Cepstral Coefficients) et ont trouvé des performances améliorées (Acc=87.5%) par rapport à leurs travaux antérieurs avec les MFCC.

Au final ces études ayant abordé la détection de MP par le biais des MFCC ou autres paramètres cepstraux rapportent des performances allant de 70 à 99%. Il faut néanmoins prendre en compte deux considérations : la première étant que les meilleures performances rapportées sont, pour la plupart, issues de méthodologies avec haut risque de sur-apprentissage. Le deuxième aspect est que ces classifications ne concernent pas spécifiquement le stade débutant de MP. On pourrait donc s'attendre à de moins bonnes performances si seuls les MP récemment diagnostiqués étaient pris en compte dans ces analyses.

3.4 Télédiagnostic de MP

Depuis quelques années, certaines études ont exploré la possibilité d'un télédiagnostic de MP (non spécifiquement débutant). Des enregistrements vocaux ont été réalisés avec des applications smartphones ou tablettes de haute qualité (taux d'échantillonnage de 44 ou 48 kHz) et envoyés dans un second temps à un serveur distant pour être analysés [Zhang et al., 2018, Benba et al., 2016b, Rusz et al., 2018, Vaiciukynas et al., 2017, Zhang, 2017, Sakar et al., 2017]. Les enregistrements audio étaient parfois combinés à d'autres modalités, telles que l'analyse des mouvements pour détecter la maladie [Arora et al., 2015, Brunato et al., 2013].

Certaines autres études ont exploré l'effet de la transmission de la voix par voie téléphonique sur la détection de MP (ou d'autres pathologies de la voix), en le simulant à partir d'enregistrements de haute qualité [Wu et al., 2018, Tsanas et al., 2012a, Vásquez-Correa et al., 2017b, Fraile et al., 2009a].
Certains projets ont commencé à acquérir des enregistrements vocaux de patients MP (non spécifiquement débutants et sans vérification par un médecin) et de sujets sains dans le but d'étudier la détection de MP via des données issues du réseau téléphonique, comme le projet *Parkinson Voice Initiative* (http ://www.parkinsonsvoice.org).

A notre connaissance les travaux que nous avons présentés en 2019 à la conférence Interspeech [Jeancolas et al., 2019a] sont les premiers à avoir traité la détection de MP via des enregistrements téléphoniques réels (issus du réseau téléphonique).

Chapitre 4

Constitution de nos bases de données

Comme expliqué dans l'introduction, ma thèse s'inscrit dans le cadre du protocole ICEBERG (cf. section 1.2). A ce titre j'ai pu enregistrer des sujets MP idiopathiques débutants (moins de 4 ans d'évolution), des sujets iRBD, et des sujets sains.

Les critères d'inclusion des sujets étaient les suivants :

Critères d'inclusion MP :

- MP idiopathique d'après les critères de l'UKPDSBB
- Diagnostic datant de moins de 4 ans
- DatScan positif (présence d'un déficit en transporteur dopaminergique)
- Absence de syndrome parkinsonien atypique : atrophie multisystématisée (MSA), paralysie supranucléaire progressive (PSP), démence à corps de Lévy (DCL)
- Absence de syndrome parkinsonien dû aux neuroleptique ou au MPTP

Critère d'inclusion iRBD :

- Troubles isolés du comportement durant la phase de sommeil paradoxal
- Absence de symptôme parkinsonien
- Examen neurologique normal

Critère d'inclusion sujets sains :

- Avoir entre 40 et 70 ans
- Examen neurologique normal

Les patients MP et iRBD ont été recrutés par des neurologues de l'Institut du Cerveau et de la Moelle (ICM) et de l'hôpital Pitié Salpêtrière. Les sujets sains ont répondu, pour la majeure partie, à une offre du RISC (relai d'expériences en sciences cognitives). J'ai pu enregistrer au moins une fois la quasi-totalité de ces sujets à l'hôpital, et par téléphone la majorité d'entre eux.

Comme un nombre restreint de sujets sains avaient été recrutés dans le cadre de l'étude ICEBERG au début de ma thèse, j'ai recruté 51 autres sujets sains de la même tranche d'âge, de mon entourage, que j'ai enregistrés chez eux ou à leur travail avec le même matériel d'enregistrement. Parmi ces derniers, certains d'entre eux ont également effectué les enregistrements téléphoniques. Les bases de données constituées et utilisées sont détaillées dans les sections suivantes.

Deux informations supplémentaires peuvent être apportées pour préciser le statut des sujets : la présence d'un syndrome RBD chez les MP et une hypothèse de conversion MP chez les iRBD.

Pour savoir si un sujet MP a en plus un syndrome RBD les conditions suivantes, doivent être vérifiées :

- résultat du questionnaire REM sleep Behaviour Disorder Questionnaire - Hong Kong (RBDQ-HK) [Li et al., 2010] > 13

- et activité tonique du menton en sommeil paradoxal > 18% du temps de sommeil paradoxal d'après la polysomnographie

- ou présence d'un comportement onirique sur la vidéo de polysomnographie

D'après ces critères, 35% des sujets MP que j'ai enregistrés présentent un syndrome RBD.

Pour la conversion des sujets iRBD en MP, la question est plus compliquée car l'entrée dans la maladie de Parkinson est assez floue et ne fait pas l'objet d'une définition unanime au sein de la communauté scientifique. Néanmoins nous pouvons observer l'apparition de symptômes moteurs chez certains des iRBD de notre base. Nous appellerons moteurs positifs les iRBD dont le score MDS-UPDRS III est supérieur à 14, et moteurs négatifs ceux donc le score est inférieur.

D'après ce critère, et après exclusion de deux sujets iRBD qui seraient en train d'évoluer vers une DCL, 44% des sujets iRBD que j'ai enregistrés peuvent être considérés comme moteurs positifs.

4.1 Enregistrements en condition de laboratoire

4.1.1 Participants

Les chiffres relatifs aux bases de données voix sont ceux correspondant à l'état des enregistrements en février 2019, moment auquel j'ai gelé la base pour faire les analyses finales. A cette époque nous avions enregistré à l'hôpital de la Pitié Salpêtrière 227 sujets de l'étude ICEBERG. Comme les sujets viennent une fois par an dans le cadre du protocole, j'ai pu enregistrer certains sujets (environ la moitié) 2 fois et une dizaine 3 fois, mais seule la première session a été utilisée pour les analyses de ma thèse. Les autres sessions serviront pour une étude longitudinale quand le protocole sera terminé.

En plus des 227 sujets ICEBERG enregistrés à l'hôpital, j'ai enregistré 51 autres sujets sains de mon entourage avec le même matériel d'enregistrement. Les sujets ICEBERG ont été enregistrés dans des box de consultations et les sujets sains supplémentaires ont été enregistrés chez eux ou à leur travail.

Nous avons dû retirer certains sujets pour les analyses, comme des sujets qui se sont révélés après coup incompatibles avec les critères d'inclusion, des sujets ayant des pathologies du langage autres que dues à MP (bègue..). Certains enregistrements n'ont également pas pu être pris en compte à cause de problèmes techniques lors de l'enregistrement ou de non-respect des consignes. Les sujets gardés pour les analyses sont au nombre de **252**, composés de **115 MP** (74 hommes et 41 femmes), de **46 iRBD** (41 hommes et 5 femmes) et de **91 sains** (48 hommes et 43 femmes). Parmi les 91 sains, 45 sont des sujets du protocole ICEBERG (23 hommes et 22 femmes) et 46 sont des sujets que j'ai enregistrés en plus (25 hommes et 21 femmes). Les statistiques concernant l'âge, les scores moteurs d'UPDRS III OFF et les niveaux sur l'échelle de Hoehn et Yahr sont détaillés Figure 4.1. Au moment des enregistrements la grande majorité des patients MP étaient traités pharmacologiquement, seuls 3 patients ne suivaient pas de traitement. Nous avons enregistré les patients traités en état de ON "étendu" à savoir moins de 12h après leur prise médicamenteuse du matin.

4.1.2 Tâches vocales

Le premier travail a été de choisir les tâches vocales les plus pertinentes à faire faire aux sujets afin de mettre le plus possible en valeur les différences vocales entre les sujets sains et

	Nombre	Âge		UPDRS III off		Hoehn & Yahr	
		moy	SD	moy	SD	moy	SD
MP	**115**	**63,77**	**9,27**	**32,49**	**6,97**	**2,01**	**0,09**
F	41	63,89	9,32	29,56	5,84	2,00	0,00
M	74	63,70	9,31	34,14	7,05	2,01	0,12
RBD	**46**	**68,58**	**6,20**	**14,15**	**8,28**	**0,74**	**0,95**
F	5	71,64	4,47	19,40	7,02	0,80	1,10
M	41	68,21	6,32	13,51	8,27	0,74	0,95
sain	**91**	**59,11**	**9,98**	**4,77**	**3,54**	**0,04**	**0,30**
F	43	59,32	9,25	4,91	3,42	0,09	0,43
M	48	58,93	10,68	4,64	3,74	0,00	0,00
Total	**252**	**62,97**	**9,65**	**22,38**	**13,60**	**1,31**	**0,96**

FIGURE 4.1 – Base de données, utilisée pour l'analyse, des sujets enregistrés en condition de laboratoire, avec le microphone professionnel et le microphone interne de l'ordinateur.

les sujets Parkinsoniens ou pré-Parkinsoniens. A cette fin j'ai fait un état de l'art [Jeancolas et al., 2016] sur l'analyse de la voix dans la maladie de Parkinson que j'ai présenté lors d'une conférence. La voyelle soutenue /a/ semble suffire à elle seule pour détecter la maladie de Parkinson à un stade relativement avancé (elle a été beaucoup utilisée dans ce cadre). Cependant pour les stades débutants, d'autres types de tâches semblent plus pertinents. Le Tableau 2.2 regroupe les tâches les plus pertinentes d'après la littérature dans la détection précoce de Parkinson et les paramètres acoustiques discriminants associés. Cet état de l'art m'a permis de sélectionner un ensemble de tâches à proposer aux participants :

- **Voyelle soutenue** : maintenir le son /a/ le plus longtemps possible en un seul souffle. Cette tâche met en évidence les problèmes de phonation (changement de timbre, et difficulté à maintenir une hauteur (F0) et une intensité constantes). On a choisi la voyelle /a/ car des études ont montré qu'elle permettait un meilleur taux de détection de la maladie de Parkinson que les autres voyelles [Orozco-Arroyave et al., 2014a].

- **Glissando** : prononcer le son /a/ à la manière d'une sirène (du grave à l'aigu puis redescente vers le grave). Nous avons sélectionné cette tâche, introduite très récemment dans des études de détection de Parkinson [Orozco-Arroyave et al., 2014a] pour tester les difficultés de continuité du mouvement et d'amplitude du spectre vocal.

- **Diadococinésie (DDK)** : répétitions rapides de syllabes en un seul souffle (/pa/, /pou/, /kou/, /poupa/, /pakou/, /pataka/, /badaga/, /patikou/, /pabikou/, /padikou/). Ces répétitions rapides font apparaître les problèmes d'articulation qui ont surtout lieu pendant la prononciation de consonnes occlusives (cf. section 2.3.2). Les voyelles choisies entre ces consonnes sont celles qui forment le triangle vocalique.

- **Monologue** : raconter sa journée pendant une minute. Cette tâche permet d'étudier la diminution de la prosodie et les problèmes d'articulations (de consonnes comme de voyelles). Cette tâche permet également d'avoir un jeu de données texte-indépendant.

- **Lectures** : un petit texte, un dialogue et 2 phrases courtes (1 question et 1 exclamation) à lire. La lecture permet d'étudier les difficultés de prosodie et d'articulation mais sans variation due au contenu linguistique. Le texte contient tous les phonèmes de la langue française. Le dialogue et les phrases courtes sont à valeur émotionnelle de façon à impliquer une prosodie prononcée. Ci-dessous l'ensemble des phrases et textes que les sujets ont eu à lire :

- texte :
"Au loin un gosse trouve, dans la belle nuit complice, une merveilleuse et fraîche jeune campagne. Il n'a pas plus de dix ans et semble venir de très loin. Comment il en est arrivé là, ça l'histoire ne le dit pas."

- dialogue :
— Tu as eu des nouvelles de Ludivine récemment ? Elle ne répond plus à mes messages depuis quelques temps.
— Je l'ai aperçue par hasard au parc hier. Tu ne devineras JAMAIS ce qu'elle faisait !
— Vas-y raconte !
— Elle courait autour du stade à CLOCHE-PIED et avec un BANDEAU sur les yeux !
— Ha la la ! Comment c'est possible de ne pas avoir peur du ridicule à ce point ?
— À mon avis elle aime juste bien se faire remarquer.

- phrases courtes :
- Tu as appris la nouvelle ?
- C'est pas possible !

- **Répétitions de phrases courtes** : 2 questions et 2 exclamations à répéter. Nous avons ajouté cette tâche pour comparer les problèmes de prosodie et de répétition lors de la lecture et lors d'une répétition. Une des deux questions et une des deux exclamations sont les mêmes que celles de la lecture. Dans le but que la prononciation de ces phrases lors de la tâche de lecture ne soit pas influencée par l'écoute de ces mêmes phrases lors de la tâche de répétition, notre algorithme fait apparaître les 2 phrases communes d'abord dans la tâche de lecture. Les répétitions de phrases courtes sont d'autant plus importantes pour l'étude téléphonique que nous n'y proposons pas de tâche de lecture pour des raisons pratiques. Ci-dessous les textes des phrases que les sujets ont eu à répéter :

  - "Tu as appris la nouvelle ?"
  - "C'est pas possible !"
  - "Tu sais ce qu'il est devenu ?"
  - "Il n'aurait jamais dû faire ça !"

- **Rythme** : Répéter lentement les syllabes /pa/, /kou/ et /pa kou/ en essayant de respecter le rythme de l'exemple (soit une syllabe par seconde) et ce pendant 30s. L'utilité de cette tâche est qu'elle met en valeur la difficulté qu'ont les sujets Parkinsoniens de maintenir un rythme constant [Skodda et al., 2013].

- **Silence** : Un silence de 5s est enregistré, pendant lequel on demande au sujet de respirer normalement. Cet enregistrement permet de capturer le bruit de fond.

Toutes ces tâches ont été faites une fois par session sauf la répétition rapide de /pataka/, la répétition lente de /pa kou/, la voyelle soutenue et le glissando qui ont été effectuées 2 fois.

A raison de 6 min d'enregistrements en moyenne par sujet, la quantité de données paroles totale disponibles pour l'analyse de cette base de données est d'environ 25 heures (11.5 h de MP, 4.5 h de iRBD et 9 h de sains). Le détail de la quantité de parole par type de tâche est présenté Figure 4.2.

4.1.3 Acquisitions

Pour que les enregistrements se fassent de manière automatique et dans les mêmes conditions à chaque fois, j'ai développé une interface utilisateur sur Matlab, qui affiche les énoncés des

	Nombre de sujets	aaa (12s)	glissando (8s)	DDK (1min30)	dont pataka (20s)	Monologue (1min)	Lecture (1min)	Répétitions (10s)	Rythme (2min)	Total (6min)
MP	**115**	**23**	**15,3**	**172,5**	**38,3**	**115,0**	**115,0**	**19,2**	**230**	**690**
F	41	8,2	5,5	61,5	13,7	41,0	41,0	6,8	82	246
M	74	14,8	9,9	111,0	24,7	74,0	74,0	12,3	148	444
RBD	**46**	**9,2**	**6,1**	**69,0**	**15,3**	**46,0**	**46,0**	**7,7**	**92**	**276**
F	5	1	0,7	7,5	1,7	5,0	5,0	0,8	10	30
M	41	8,2	5,5	61,5	13,7	41,0	41,0	6,8	82	246
sain	**91**	**18,2**	**12,1**	**136,5**	**30,3**	**91,0**	**91,0**	**15,2**	**182**	**546**
F	43	8,6	5,7	64,5	14,3	43,0	43,0	7,2	86	258
M	48	9,6	6,4	72	16	48	48	8	96	288
Total	**252**	**50,4**	**33,6**	**378**	**84**	**252**	**252**	**42**	**504**	**1512**

FIGURE 4.2 – Quantité de données paroles disponibles pour l'analyse de la base de données enregistrées en condition de laboratoire. Les quantités sont exprimées par type de tâche et en minute. En entête figure la durée moyenne de chaque type de tâche.

tâches, permet d'écouter des exemples et enregistre la voix des sujets. Les tâches apparaissent dans un ordre aléatoire pour s'affranchir d'un éventuel biais dû à leur ordre. Un opérateur (moi-même ou depuis octobre 2018 des collègues du centre d'investigation clinique de l'ICM) est au côté des sujets pendant la session d'enregistrement pour surveiller que les tâches soient correctement effectuées et répondre à leurs questions. Les tâches peuvent être refaites quand cela est nécessaire. La session d'enregistrement dure 15-20 min (consignes comprises) et commence par 3 min de questions. Nous demandons aux sujets leur langue maternelle, s'ils prennent un traitement pour la maladie de Parkinson et si oui de quand date leur dernière prise, s'ils ont déjà vu un orthophoniste et pour quelle raison, si eux ou leur entourage ont remarqué des changements dans leur voix (et quel type de changement) et s'ils ont en ce moment une angine, rhinite ou autre qui pourrait modifier leur voix. Les autres détails les concernant (comme la date de naissance, les différents traitements qu'ils prennent, le fait de fumer ou non...) leur sont demandés lors d'un entretien avec un neurologue de l'équipe.

Les enregistrements ont été effectués avec deux microphones fonctionnant en parallèle : un microphone de qualité professionnelle, et le microphone interne de l'ordinateur servant aux acquisitions, dans le but d'évaluer l'importance de la qualité des microphones sur nos analyses.

4.1.3.1 Enregistrements avec le microphone professionnel

Pour les enregistrements de très bonne qualité, j'ai choisi d'utiliser un microphone omni-directionnel à électret (cf. section 2.1.2) afin d'avoir le moins de déformations possibles de la voix lors de l'enregistrement (tout en respectant un coût abordable). Je l'ai pris en serre tête de façon à ce qu'on puisse analyser les variations d'intensité au cours d'une même tâche (ce qui nécessite d'avoir une distance constante entre la bouche et le microphone). J'ai choisi la modèle *Beyerdynamics Opus 55 mk ii* car il donnait une réponse fréquentielle particulièrement plate. Nous le plaçons à 5-10 cm de la bouche des sujets, sur le côté, de façon à éviter les perturbations dues aux flux d'air émis par le locuteur. Ce microphone est connecté à une carte son externe professionnelle (*Scarlett 2i2, Focusrite*), qui lui fournit une alimentation fantôme et procure une préamplification. La voix est échantillonnée à 96kHz et l'étendue spectrale du microphone est de [50Hz, 20kHz]. L'audio est encodée au format PCM S24 LE, ce qui signifie d'après la section 2.1.2, que les amplitudes sont codées sur 24 bits, de manière signée et en little endian.

4.1.3.2 Enregistrements avec le microphone de l'ordinateur

De manière à contrôler la nécessité d'une telle qualité d'enregistrement, toutes les tâches ont été enregistrées en même temps avec le microphone interne de l'ordinateur (*MacBook Air mi-2012, OS X Yosemite*) à 96kHz. Le format d'encodage est identique, soit PCM S24 LE. La

fonction de réduction de bruit est activée par défaut, nous avons choisi de la maintenir activée. Le microphone étant dans l'ordinateur, la distance moyenne avec la bouche des participants est d'environ 50cm.

4.2 Enregistrements par téléphone

Dans le but d'évaluer si la détection précoce de MP par l'analyse de la voix est possible via des enregistrements téléphoniques, nous avons proposé aux participants ICEBERG et à quelques-uns des sujets sains supplémentaires, d'effectuer des tâches vocales du même type que précédemment mais cette fois-ci par téléphone. Pour compenser la qualité réduite des enregistrements téléphoniques (par rapport à ceux faits en laboratoire), nous avons souhaité augmenter la quantité des enregistrements en proposant aux sujets de les faire une fois par mois. De plus le fait d'enregistrer les sujets régulièrement permettra de mieux évaluer l'évolution de la voix due à la progression de MP. Tout comme pour les enregistrements au laboratoire, nous avons figé la base de données des enregistrements téléphoniques en février 2019 pour les analyses.

4.2.1 Participants

La majorité des sujets enregistrés en condition de laboratoire ont effectué un ou plusieurs enregistrements par téléphone. Au total 210 personnes ont participé à ces enregistrements. **200 sujets** ont été gardés pour les analyses, dont **101 MP** (63 hommes et 38 femmes), **38 iRBD** (36 hommes et 2 femmes) et **61 sains** (36 hommes et 25 femmes). Parmi les sujets sains, 46 sont des sujets du protocole ICEBERG et 15 font partie de la base additionnelle.

Les statistiques concernant l'âge, les scores moteurs d'UPDRS III off et les niveaux sur l'échelle de Hoehn et Yahr sont détaillés Figure 4.3.

	Nombre	Âge		UPDRS III off		Hoehn & Yahr	
		moy	SD	moy	SD	moy	SD
MP	**101**	**63,54**	**9,05**	**32,40**	**6,97**	**2,01**	**0,10**
F	38	63,34	9,30	29,50	6,06	2,00	0,00
M	63	63,66	8,98	34,18	6,94	2,02	0,13
RBD	**38**	**68,13**	**6,23**	**12,71**	**8,23**	**0,63**	**0,91**
F	2	67,80	5,27	12,50	4,95	0,00	0,00
M	36	68,15	6,34	12,72	8,42	0,67	0,92
sain	**61**	**62,59**	**8,53**	**4,89**	**3,50**	**0,04**	**0,29**
F	25	61,84	7,36	5,26	3,56	0,11	0,46
M	36	63,11	9,32	4,62	3,51	0,00	0,00
Total	**200**	**64,12**	**8,62**	**21,55**	**13,89**	**1,25**	**0,97**

FIGURE 4.3 – Base de données, utilisée pour l'analyse, des sujets enregistrés au téléphone.

Ces sujets ont effectué entre 1 à 13 sessions d'enregistrement, suivant leur moment d'inclusion, et le souhait de certains de ne pas poursuivre ce protocole d'enregistrement, aboutissant à une moyenne de 5 sessions par sujet, ce qui fait un total d'environ 1000 sessions d'enregistrement téléphoniques utilisables.

4.2.2 Tâches vocales

Les tâches vocales sont du même type que celles effectuées à l'hôpital, mais en un peu plus court (les sessions durent 12 min au lieu de 15-20 min) et sans tâches de lecture. Pour des raisons pratique, nous avons souhaité que toutes les consignes soient orales, sans envoi au préalable de textes à lire. Les tâches sont toutes effectuées une seule fois par session (pour éviter d'allonger la durée des sessions). Ci-dessous est présenté le détail des tâches vocales :

- **Voyelle soutenue** : maintenir le son /a/ le plus longtemps possible en un seul souffle.

- **Glissando** : prononcer le son /a/ à la manière d'une sirène (du grave à l'aigu puis redescente vers le grave).

- **Diadococinésie (DDK)** : répétitions rapides de syllabes en un seul souffle (/pa/, /pou/, /kou/ /poupa/, /pakou/, /pataka/).

- **Répétitions de phrases courtes** : 8 phrases courtes à répéter, dont 3 questions et 3 exclamations et 2 phrases déclaratives. Parmi ces phrases, 4 sont identiques à celles répétées lors de la session d'enregistrement à l'hôpital, de manière à pouvoir évaluer l'influence du téléphone sur l'analyse des mêmes phrases. Ci-dessous les textes des phrases que les sujets ont eu à répéter :

  - "Tu as appris la nouvelle ?"
  - "C'est pas possible !"
  - "Tu sais ce qu'il est devenu ?"
  - "Il n'aurait jamais dû faire ça !"
  - "Tu as bien raison ! "
  - "Comment il s'appelle déjà ? "
  - "Les chiens aiment courir après les ballons. "
  - "Un carré est un rectangle particulier. "

- **Monologue** : raconter sa journée pendant une minute.

- **Rythme** : Répéter lentement les syllabes /pa/, /kou/ et /pa kou/ en essayant de respecter le rythme de l'exemple (soit une syllabe par seconde), et ce pendant 30s.

- **Citation** : La dernière tâche est une citation philosophique à répéter. Cette citation, qui change automatiquement tous les mois, a moins pour but d'être analysée que de rendre les sessions d'enregistrement mensuelles moins répétitives et rébarbatives.

A raison de 4 min d'enregistrement en moyenne par sujet et par session, et de 5 sessions en moyenne par sujet, la quantité de données paroles totale disponibles pour l'analyse de la base de données téléphoniques est d'environ 67 heures (34 h de MP, 13 h de iRBD et 20 h de sains). Le détail de la quantité de parole par type de tâche est présenté Figure 4.4.

En Annexe A figure le détail de ce que les participants entendent quand ils appellent le répondeur interactif.

4.2.3 Acquisitions

Les participants appellent une fois par mois un logiciel de téléphonie IP de type répondeur interactif qui leur fait faire automatiquement les tâches présentées ci-dessus. Ils peuvent utiliser leur téléphone fixe ou leur portable, mais doivent garder le même téléphone pendant la durée du protocole (dans la mesure du possible). Quand les participants appellent, ils sont automatiquement identifiés par leur numéro de téléphone, rentré préalablement dans le logiciel. De manière à savoir si les MP font les enregistrements en ON ou OFF, le moment de leur dernière prise médicamenteuse et le nom de leur médicament leur sont demandés par le logiciel. Les participants ont comme consigne de ne pas utiliser le mode haut-parleur du téléphone (car cela nuit trop à la qualité) et de préciser s'ils utilisent des oreillettes.

	Nombre de sujets	aaa (6s)	glissando (4s)	DDK (1min)	dont pataka (10s)	Répétitions (20s)	Monologue (1min)	Rythme (1min30)	Total (4min)
MP	**101**	**50,5**	**33,7**	**505**	**84,2**	**168,3**	**505**	**757,5**	**2020**
F	38	19,0	12,7	190	31,7	63,3	190	285,0	760
M	63	31,5	21,0	315	52,5	105,0	315	472,5	1260
RBD	**38**	**19,0**	**12,7**	**190**	**31,7**	**63,3**	**190**	**285,0**	**760**
F	2	1,0	0,7	10	1,7	3,3	10	15,0	40
M	36	18,0	12,0	180	30,0	60,0	180	270,0	720
sain	**61**	**30,5**	**20,3**	**305**	**50,8**	**101,7**	**305**	**457,5**	**1220**
F	25	12,5	8,3	125	20,8	41,7	125	187,5	500
M	36	18,0	12,0	180	30,0	60,0	180	270,0	720
Total	**200**	**100**	**66,7**	**1000**	**166,7**	**333,3**	**1000**	**1500**	**4000**

FIGURE 4.4 – Quantité de données paroles disponibles pour l'analyse de la base de données téléphoniques. Les quantités sont exprimées par type de tâche et en minute, toutes sessions confondues. En entête figure la durée moyenne de chaque type de tâches pour une session.

4.2.3.1 Mise en place du répondeur interactif

Afin que les participants puissent effectuer les enregistrements mensuels de manière automatique, nous avons mis en place un serveur vocal interactif connecté au réseau, de manière à ce que les participants puissent faire les tâches de leur téléphone en appelant un simple numéro. Nous avons opté pour le répondeur interactif IVM de l'entreprise NCH, que nous avons relié à une ligne SIP (ippi). Le déroulement de la mise en place (concernant le répondeur interactif, la ligne SIP et la connexion internet), les choix retenus et le suivi des appels sont détaillés en Annexe B.

4.2.3.2 Analyse de la chaine de transmission finale

La solution retenue pour notre répondeur téléphonique interactif fait intervenir la chaine de transmission suivante : Le participant appelle le numéro de notre ligne SIP d'ippi à partir de son téléphone fixe ou portable. Une fois la session SIP entre notre ligne ippi et notre serveur initialisée, il entend les premières consignes et peut effectuer les tâches vocales. Sa voix est transmise en analogique jusqu'au premier central téléphonique s'il appelle d'un téléphone fixe, et est transmise en numérique avec le protocole GSM jusqu'à l'antenne relai s'il utilise son téléphone portable. Dans ce dernier cas la voix est encodée en 8kHz sur 13 bits linéaires avec la bande passante de [300-3400Hz] puis compressée selon un des 9 modes possibles du codec AMR. Ensuite à partir du central téléphonique ou de l'antenne relai, la voix est transformée au format PCM -alaw, grâce au codec G.711a (cf. partie 2.1.3). La voix échantillonnée à 8kHz a une profondeur de codage de 8 bits non linéaire et une bande passante de [300-3400Hz]. Une fois arrivée au serveur ippi elle est retransformée en PCM -μ law (codec G.711μ) et transmise en VoIP à notre box internet qui transmet les données au logiciel IVM. Ce dernier transforme le G.711μ qu'il reçoit au format standard audio de PCM S16 LE, en gardant la fréquence d'échantillonnage de 8kHz. Cette chaine de transmission est illustrée Figure 4.5 .

On peut noter que même si les participants avaient appelé à partir d'un logiciel VoIP, la qualité n'aurait pas été tellement améliorée, car IVM ne supportant que le G.711 le signal aurait quand même dû être transmis en 8kHz sur 8 bits non linéaires avec une plage de fréquence de [300-3400 Hz] et aurait quand même subi la distorsion due à ce codec. Afin d'évaluer la distorsion due à ces codecs nous avons encodé un son sinusoïdale pur à la fréquence d'1kHz en G.711a et G.711μ puis nous l'avons décodé. La distorsion relative obtenue est représentée Figure 4.6. On observe que ces codecs peuvent engendrer une distorsion allant jusqu'à 5% du signal initial.

Afin d'étudier de près et de comprendre les différents problèmes de connexion et de trans-

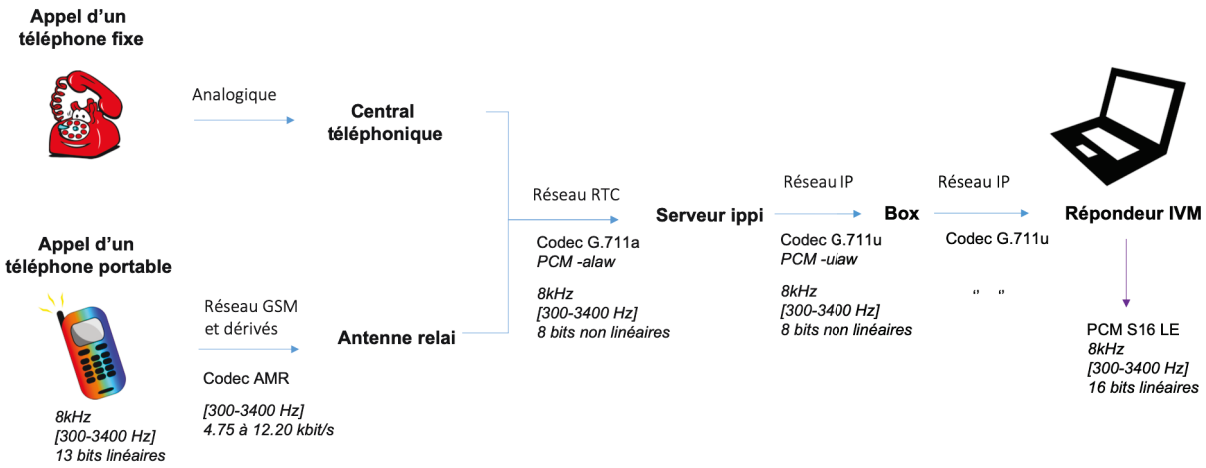


FIGURE 4.5 – Chaine de transmission de la voix des participants lors d'une session d'enregistrement téléphonique.

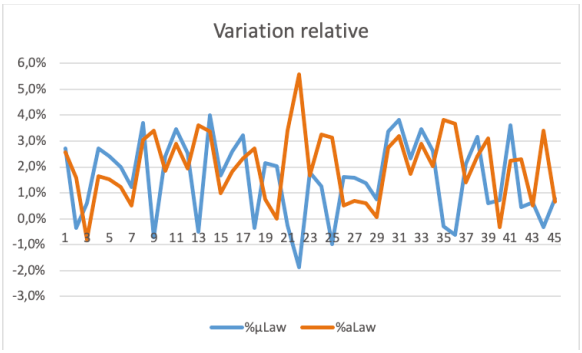


FIGURE 4.6 – Distorsion relative due au codage en G.711a et G.711μ d'un son sinusoïdal pur de fréquence 1kHz.

mission que nous avons rencontrés dans la mise en place du répondeur interactif, nous avons à plusieurs reprise enregistré et analysé, avec le logiciel WireShark, le flux réseau arrivant sur la carte Ethernet du serveur.

La première étape une fois le logiciel IVM mis en route est une étape d'enregistrement (dite *Register*) du serveur auprès de l'opérateur ippi. La trace du flux réseau correspondant à cette étape est illustrée Figure 4.7. Une demande de Register est envoyée par le serveur (adresse IP : 192.168.1.10) à l'opérateur ippi (adresse IP : 194.169.214.30). Cette demande est d'abord refusée (probablement à cause du temps de réponse de l'opérateur pour répondre OK) puis acceptée. Le Register se fait sur une session SIP et le port utilisé est le port 5060 (port standard du protocole SIP).

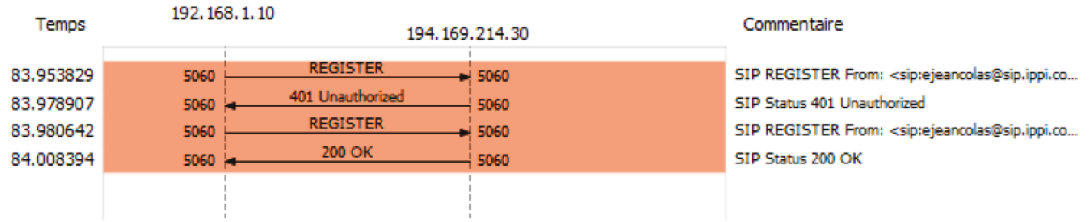


FIGURE 4.7 – Trace du flux réseau du serveur téléphonique pendant l'étape de Register.

La trace du flux correspondant à un appel vers le serveur et à l'exécution de quelques tâches

vocales, est illustrée Figure 4.8. Une fois le Register établi, lorsqu'on appelle le numéro de notre ligne ippi, ippi (adresse IP : 194.169.214.30) indique, via le protocole SIP, à notre serveur (adresse IP : 192.168.1.10) que le numéro de téléphone 33673996658 demande une connexion (INVITE). Ippi envoie également l'information qu'il accepte les codecs G.711a et G.711μ. Le serveur envoie l'information au logiciel IVM, qui attend 4 sonneries pour accepter la connexion (OK SDP) et précise qu'il accepte aussi les codecs G.711a et G.711μ, ce à quoi répond ippi par un accusé de réception (ACK). S'établit ensuite la session vocale avec le protocole RTP entre notre serveur téléphonique et un serveur RTP d'ippi (194.169.214.13). Les flux audio sont transmis en G.711μ. Le flux voix du serveur vers l'appelant se fait en un seul flux RTP, envoyant ici 6270 paquets de 20ms (les paquets de silence ne sont pas transmis). Le flux audio de l'appelant vers le serveur est entrecoupé à chaque fois que l'appelant appuie sur une touche (pour passer à la tâche suivante). On voit notamment qu'ippi donne l'information de la touche appuyée (la touche 1), ce qui signifie que l'action de décodage du code DTMF se fait bien par ippi et non par IVM, contrairement à la ligne SIP de free que nous avions testée et qui avait entraîné des problèmes de décodage DTMF (cf. Annexe B). La dernière étape est la fin de session, quand l'appelant raccroche, une commande BYE est envoyée par le serveur à ippi, via de nouveau le protocole SIP, pour clore la session.

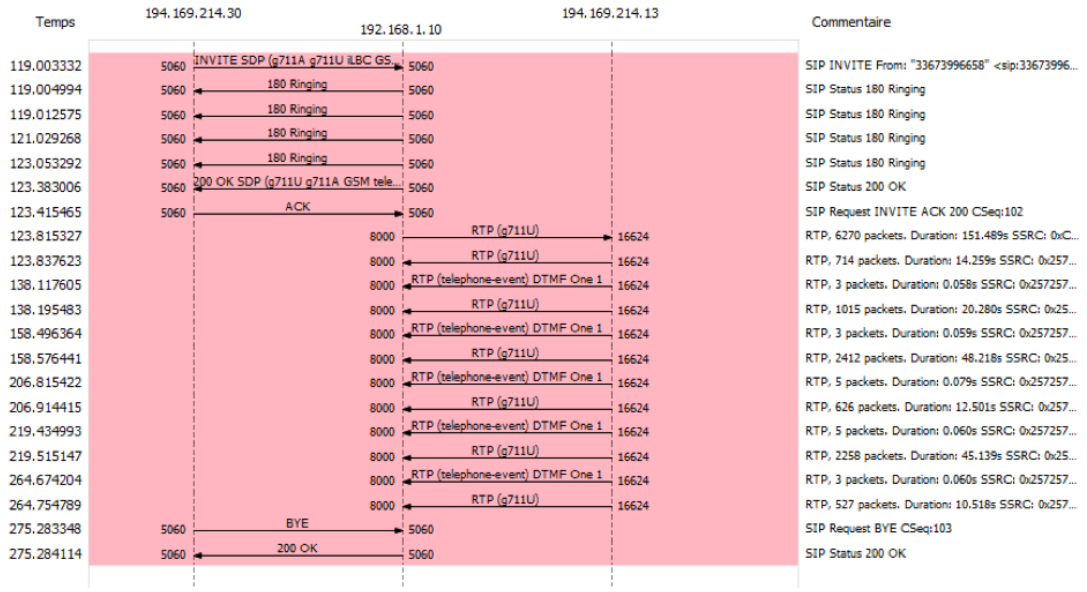


FIGURE 4.8 – Trace du flux réseau du serveur téléphonique pendant un appel d'un participant.

Une fois la session terminée, l'outil d'analyse de Wireshark nous apporte des précisions supplémentaires sur la liste des différents paquets transmis. Notamment on a accès aux numéros des paquets (attribués par le récepteur), aux numéros des séquences (attribués par l'émetteur), au nombre de paquets perdus (correspond à des numéros de séquences manquants) et au nombre d'inversions de paquets, c'est à dire quand les paquets n'arrivent pas dans le bon ordre (correspond à des numéros de séquences intervertis).

IVM a un *buffer* permettant de stocker quelques paquets avant de les "jouer" et ainsi compenser la gigue (cf. partie 2.1.3) lorsqu'elle n'est pas trop importante. Quand un paquet arrive avec trop de retard ou quand il n'arrive pas, il est considéré comme manquant et n'est pas joué. Afin de limiter la dégradation due à la coupure du signal, plusieurs techniques de masquage de paquets manquants, peuvent être utilisées pour lisser le signal [Koenig, 2011]. Nous avons voulu savoir si IVM interpolait les paquets manquants et si oui quelle technique était utilisée. Pour cela nous avons analysé une trace (flux réseau arrivant sur le serveur) avec un

paquet manquant et avons reconstitué le signal sonore en décodant le signal reçu G711u avec la table de transcodage (G711u vers PCM 16bits) de NCH pour pouvoir le comparer au signal décodé et joué par IVM. Les deux signaux sont représentés Figure 4.9, en bleu pour le signal G711u que nous avons décodé, et en orange le signal joué par IVM. On peut constater que les signaux se superposent sauf au moment du paquet manquant (20ms) et à l'arrivée du paquet suivant. IVM effectue bien une interpolation au moment du paquet manquant. Le logiciel rejoue plusieurs fois le dernier tiers du paquet précédent en l'atténuant progressivement. La reprise s'effectue également de manière progressive sur un quart du paquet suivant. Un autre exemple d'interpolation d'IVM, cette fois sur 3 paquets manquants consécutifs, est représenté Figure 4.10. On retrouve la duplication d'environ un tiers du paquet précédent et l'atténuation progressive (l'atténuation atteint 90% au bout de 80ms), puis la reprise progressive du paquet suivant sur son premier quart. Ce n'est pas rare qu'il y ait un ou plusieurs paquets manquants lors d'une session d'enregistrement d'un participant, même si l'idéal serait de ne pas en avoir, l'interpolation effectuée par notre logiciel de téléphonie limite la dégradation de la voix et de son analyse.

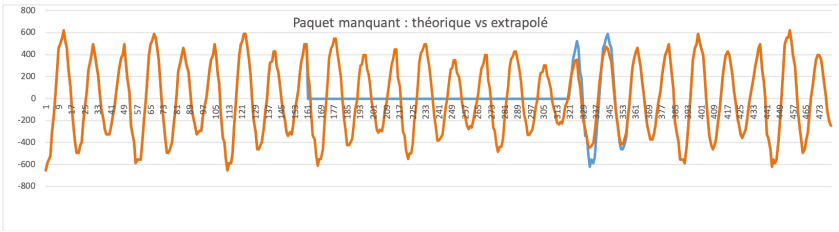


FIGURE 4.9 – Interpolation d'un paquet manquant par IVM. En bleu est tracé le signal G711u reçu par le serveur et décodé d'après la table de transcodage de NCH, et en orange le même signal décodé puis interpolé par IVM.

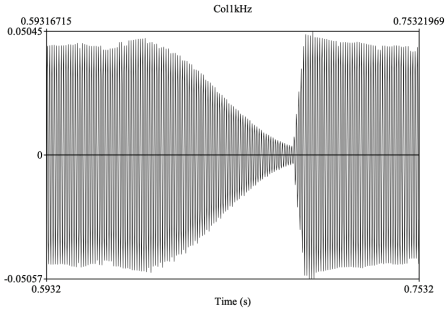


FIGURE 4.10 – Interpolation de 3 paquets manquants consécutifs par IVM

En analysant de plus près le contenu d'un paquet avec Wireshark on constate qu'il est composé de 160 octets de données paroles et 54 octets d'informations supplémentaires (adresses IP, protocole utilisé, numéro du paquet ...). Les données paroles étant encodées avec le codec G.711μ, soit 8 bits par valeur avec une fréquence de 8kHz, 160 octets correspondent alors à une durée parole de 20ms. Les paquets transmettent donc bien des morceaux de voix de 20ms. Le débit correspondant à cette transmission voix est de 64kbit/s. En rajoutant les 54 octets supplémentaires par paquet, on atteint 85kbit/s, on retrouve l'ordre de grandeur du débit moyen constaté sur le serveur lors d'une session d'enregistrement (cf. Annexe B).

4.3 Validations et prétraitements

4.3.1 Enregistrements en condition de laboratoire

Afin de constituer une base de données voix de la meilleure qualité possible, nous avons écouté et validé "à la main" au moins une session enregistrée en condition laboratoire par sujet, soit 280 sessions, donc à raison de 29 tâches par session, cela faisait 8120 fichiers audios. La validation a consisté en une annotation avec Praat des fichiers sons, permettant d'identifier les anomalies (erreurs d'exécution de la part du sujet et commentaire, défauts techniques, saturations, rire, bruit extérieur ...) et de délimiter les portions des tâches de bonne qualité.

Nous avons écarté pour l'analyse, les sessions avec des anomalies majeures (absence de son due à une mauvaise connexion du microphone, fou rire sur plusieurs tâches, saturations importantes, difficultés à parler dues à des pathologies du langage (quelques sujets bègues...). Dans le cas où une autre session avait été enregistrée l'année suivante (et correctement effectuée), cette dernière était prise en compte. En plus des sujets écartés suite aux annotations, quelques sujets supplémentaires n'ont pas été pris en compte dans l'analyse car s'étant avérés incompatibles avec les critères d'inclusion du protocole, résultant en un total de 252 sujets analysables (cf. section 4.1.1). Pour ces sujets, l'analyse a été effectuée sur la portion de chaque tâche identifiée comme de bonne qualité. Seul le début ou la fin de la tâche pouvait être coupée (afin d'éviter des discontinuités au milieu de la tâche) supprimant ainsi les bruits de respiration pour les tâches en apnée, les clics de souris et les commentaires que faisaient certains sujets en début ou fin de tâche. Les petites anomalies pouvant survenir au milieu de la portion gardée étaient répertoriées pour information mais non supprimées.

Nous aurions pu nous passer de l'étape de validation et garder la même qualité d'enregistrement si nous avions demandé aux sujets de refaire la tâche lors du constat d'une anomalie de type rire, commentaire, gros bruit extérieur, saturation (visible par un changement de couleur d'une diode sur la carte son)... Or nous faisions déjà recommencer les sujets lorsqu'ils ne respectaient pas la consigne, et la quantité d'autres tests qu'ils avaient à faire dans le cadre du protocole ICEBERG dans la même journée nous donnait une contrainte de durée de session à respecter. Si cela avait été possible, il aurait fallu ensuite détecter de façon automatique les fichiers sans son, couper de façon automatique les fins de tâches pour retirer les clics de souris, et les débuts des tâches en apnée pour retirer les profondes inspirations, ce que nous avons fait pour les enregistrements téléphoniques.

4.3.2 Enregistrements téléphoniques

Les enregistrements téléphoniques étant plus nombreux, environ 1000 sessions, donc à raison de 20 tâches par session, 20 000 fichiers audios, ce n'était pas envisageable de tous les écouter et les annoter à la main. Du coup nous avons automatisé la détection de fichiers manquants et de fichiers sans son. Nous avons fixé les durées à couper au début et à la fin des tâches après écoute de plusieurs enregistrements. Nous avons coupé les 10 premières ms et les 40 dernières ms de tous les fichiers pour enlever les bips des touches. En plus nous avons coupé les 10 dernières secondes du monologue et les 5 dernières secondes des tâches de rythme pour enlever les bips des touches de ceux qui essaient de taper 1 alors que les tâches ne sont pas finies et s'arrêtent toute seules (contrairement aux autres). Pour les tâches en apnée à tenir le plus longtemps possible, nous avons retiré la 1^{ere} seconde et les 3 dernières secondes pour enlever les inspirations et expirations profondes.

4.4 Constitution de bases de données supplémentaires non analysées dans ma thèse

4.4.1 Visage

L'hypomimie (la réduction des mouvements du visages) est un autre symptôme concernant quasiment tous les patients atteints de la Maladie de Parkinson [Hoehn and Yahr, 1967] [Jankovic, 2003]. Il est une manifestation de la bradykinésie et se manifeste par une diminution de l'expression faciale. Il est intéressant de pouvoir étudier par la suite ce symptôme en plus de la voix car il est également connu pour apparaître tôt dans cette maladie, et est présent chez les sujets iRBD [Postuma et al., 2012] . De manière à pouvoir étudier ultérieurement cette hypomimie, nous proposons depuis peu aux sujets ICEBERG d'enregistrer, s'ils le souhaitent, leur visage pendant les tâches vocales. Une webcam avec encodage et compression intégrée, Logitech C922 Pro Stream Webcam, est branchée sur l'ordinateur servant à l'acquisition de la voix. Les réglages que nous avons retenus sont un taux d'images par seconde de 24 fps (*frames per second*), une résolution de 1920x1080 pixels, le codec H264 et une fréquence d'échantillonnage de 44100Hz pour le son. La webcam enregistre toute la session et une écriture des temps de début et de fin de chaque tâche se fait de manière automatique dans un fichier texte, de manière à pouvoir séparer facilement les tâches par la suite. On a choisi de ne pas arrêter les enregistrements à la fin de chaque tâche, comme on le fait pour le son avec le microphone professionnel et le microphone de l'ordinateur, car la webcam met entre 0 et 2.5s pour se mettre en route à chaque fois, rallongeant alors la durée de la session d'enregistrement.

4.4.2 IRMf

De manière à pouvoir étudier ultérieurement les altérations des réseaux neuronaux associés à la parole au début de la maladie de Parkinson, nous avons demandé aux sujets du protocole ICEBERG d'effectuer des tâches vocales pendant une acquisition d'IRMf de 7 min.

Les sujets ont eu à effectuer une répétition rapide de syllabes /pa/, /kou/ et /pakou/ à voix haute et dans leur tête. Les tâches sont effectuées tant que la consigne est affichée (ce qui dure environ 16s). Les sujets ont le droit de reprendre leur souffle autant qu'ils veulent pendant les tâches.

Les tâches sont effectuées 4 fois chacune dans un ordre aléatoire et sont séparées d'une croix noire à fixer pendant 8s. Un symbole de haut-parleur ou haut-parleur barré (quand la tâche doit être effectuée silencieusement) apparaît 3s avant les syllabes à prononcer, afin d'étudier un possible effet de la maladie lors de la phase de préparation motrice [Arnold et al., 2014]. L'intérêt d'effectuer également les tâches silencieusement est de pouvoir dans l'analyse séparer les effets de la maladie sur la fin du processus moteur (mouvement des cordes vocales et des muscles d'articulation) des effets sur les étapes plus en amont du processus moteur. L'effet de la complexité des tâches pourra être étudiée (/pakou/ étant supposé plus complexe que /pa/ et /kou/).

Un jitter est appliqué au début de chaque tâche, pour éviter les effets d'anticipation.

Nous avons choisi, comme tâches vocales, des tâches de répétitions rapides car elles ne font pas intervenir de processus cognitifs linguistiques, mais juste des processus de vocalisation et d'articulation. Ceci nous permettra de ne pas confondre ces effets lors de l'interprétation des IRMf. De plus des études ont montré que l'articulation rapide était particulièrement impactée chez les iRBD [Rusz et al., 2015]. Nous avons choisi de ne pas demander aux sujets d'effectuer les tâches en apnée de manière à éviter d'avoir un mouvement de tête important en début et fin de tâche dû à une inspiration profonde. En effet ce mouvement aurait été corrélé avec les tâches et donc aurait créé un effet difficile à isoler dans l'analyse.

Nous avons effectué des IRM 3T du cerveau entier (avec cervelet) avec un multibande (3), 54 coupes transversales, des voxels de 2.5mm iso, un TR de 1405ms, un TE de 30ms. Les consignes apparaissaient sur un écran et une vidéo expliquant ces dernières était montrée aux participants juste avant la séquence. La voix des sujets a été enregistrée pendant l'IRMf à l'aide d'un micro IRM compatible. Nous avons programmé l'interface d'acquisition avec Matlab.

A ce jour (août 2019) 109 sujets ont effectué cette séquence vocale d'IRM, dont 56 MP (30 hommes et 26 femmes), 17 iRBD hommes, 36 sains (19 hommes, 17 femmes). Parmi ces sujets, 17 ne sont pas exploitables, pour cause de mauvaise exécution des tâches vocales, de problèmes techniques ou de sorties d'étude en raison d'incompatibilité avec les critères d'inclusion. Ce qui fait 92 sujets analysables dont 48 MP, 16 iRBD et 28 sains.

Chapitre 5

Classification MP vs sain avec la méthode MFCC-GMM

La première méthode de classification que nous avons choisi d'adapter pour la détection de MP est une méthode classique en reconnaissance du locuteur, utilisant des GMM pour décrire la distribution de paramètres cepstraux. Cette méthode a l'avantage de nécessiter peu de données et d'avoir un faible coût computationnel.

5.1 Méthode MFCC-GMM

5.1.1 Analyse Préliminaire

Nous avons effectué une analyse préliminaire, présentée à la conférence ATSIP [Jeancolas et al., 2017], à partir des participants que nous avions pu enregistrer avec le microphone professionnel à ce moment-là (constituant un sous-groupe de la base de données actuelle). Nous avons utilisé une méthode de classification simple, inspirée de ce qui se fait en reconnaissance du locuteur, à l'aide de la toolbox Voicebox de Matlab. Nous avons considéré les hommes et les femmes séparément car les différences au niveau des MFCC dues au genre diminuent les performances de classification en reconnaissance du locuteur. Cela a également été montré dans la détection de pathologies vocales [Fraile et al., 2009b].

La méthode consista à extraire 12 MFCC calculés sur des fenêtres temporelles de Hamming de 20ms, toutes les 10ms. 34 filtres MEL triangulaires ont été utilisés, allant de 0 à 48kHz. Nous avons ensuite créé 2 modèles GMM de dimension 12 pour modéliser les distributions de MFCC obtenus dans le groupe MP et dans le groupe sain, et ce avec seulement des sujets hommes d'un côté et femmes de l'autre. Des matrices diagonales de covariance ont été utilisées dans l'algorithme EM. Enfin un calcul de vraisemblance était effectué sur les MFCC des sujets tests par rapport aux deux modèles correspondant à leur genre, la vraisemblance la plus grande donnant le résultat de classification. Une validation croisée de type *Leave One Subject Out* (LOSO) a été effectuée afin d'obtenir un résultat précis de la performance de classification.

Les résultats obtenus pour les différentes tâches vocales sont présentés dans [Jeancolas et al., 2017]. Il faut les considérer avec précaution car nous nous sommes rendu compte par des analyses ultérieures qu'ils étaient biaisés (les performances étaient surestimées), ceci étant dû à des conditions d'enregistrement (le lieu) pas toujours appariées entre les groupes MP et sains. En effet tous les MP ont été enregistrés dans des salles de consultations à l'hôpital de la Pitié Salpêtrière ainsi que quelques sujets sains. Les autres sujets sains (recrutés en plus du protocole ICEBERG afin d'avoir assez de sujets sains pour faire des analyses de classification) avaient été enregistrés directement chez eux ou à leur lieu de travail, avec le même matériel d'acquisition.

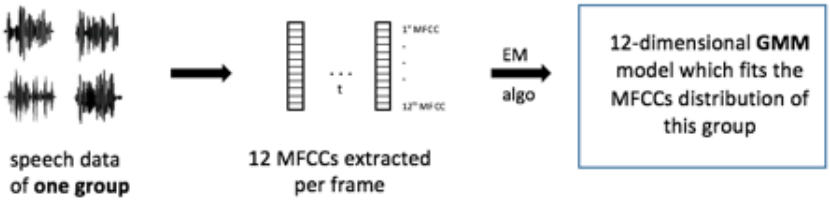


FIGURE 5.1 – Phase d'entraînement : Construction d'un modèle GMM par groupe (hommes MP, femmes MP, hommes contrôles, femmes contrôles) à partir des MFCC des sujets utilisés pour l'entraînement. EM algo : algorithme Espérance-Maximisation.

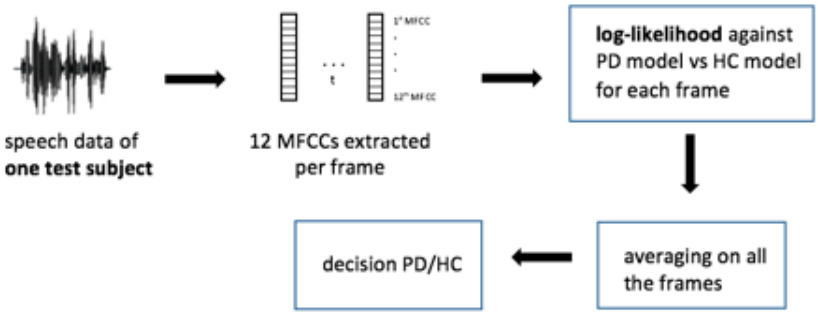


FIGURE 5.2 – Phase test : les MFCCs des sujets tests sont testés par rapport au modèle MP et control correspondant au genre. PD : Parkinson's Disease, HC : Healthy control

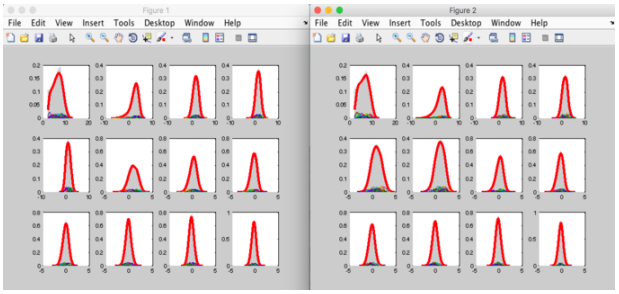


FIGURE 5.3 – Projections des GMM multidimensionnels sur les 12 MFCC. Les GMM sont entrainés sur le groupe hommes MP à gauche et hommes sains à droite

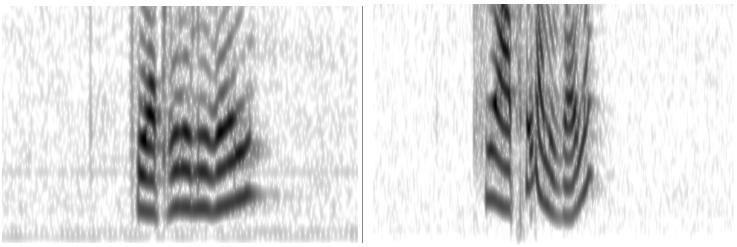
En effectuant des analyses complémentaires à l'étude préliminaire, nous nous sommes rendus compte que notre algorithme classait mieux les sujets sains enregistrés en dehors de l'hôpital que les sujets sains enregistrés à l'hôpital. Nous avons suspecté un biais provenant de la nature du bruit de fond, ce que nous avons confirmé en parvenant à classer correctement les sujets à partir de la tâche de silence. Ce qui signifiait que le lieu d'enregistrement avait un impact sur la décision de classification MP vs sain.

En examinant les spectrogrammes obtenus pendant les tâches de silence, on peut constater que ceux enregistrés à l'hôpital contenaient plus de bruit et avec souvent une ou plusieurs bandes (entre 50 et 800Hz) particulièrement marquées, dépendant du box utilisé pour les enregistrements.

Afin de supprimer ce bruit (de type *a priori* additif stationnaire), nous avons appliqué la méthode de soustraction spectrale [Boll, 1979] détaillée en partie 3.3.1.2, implémentée dans le logiciel Praat et calculée à partir de la tâche de silence de 5s.

Dans la Figure 5.4 sont représentés les spectrogrammes issus de la lecture d'une phrase, avant

et après débruitage d'un sujet enregistré à l'hôpital et d'un autre enregistré à son domicile. On constate que le bruit de fond est plus prononcé à l'hôpital avec la présence ici de bandes plus marquées autour de 50Hz et 300Hz. Après débruitage on constate que le bruit de fond a disparu pour les deux types d'environnements.



(a) Enregistrement à l'hôpital sans débruitage

(b) Enregistrement hors hôpital sans débruitage

(c) Enregistrement à l'hôpital après débruitage

(d) Enregistrement hors hôpital après débruitage

FIGURE 5.4 – Spectrogrammes de deux enregistrements effectués à l'hôpital (a) et hors hôpital, ici au domicile du sujet (b), la tache enregistrée est la lecture d'une phrase. On constate que le signal de l'hôpital est plus bruité que le signal hors hôpital avec notamment une bande à 300Hz et une bande à 50Hz. Les spectrogrammes (c) et (d) sont calculés après débruitage par soustraction cepstral.

Après ce débruitage par cette soustraction spectrale, nous avons ensuite amélioré notre méthode d'analyse sur temps courts à partir des MFCC, en améliorant l'étape d'extraction des MFCC et celle de la construction des GMM. Concernant les MFCC, nous avons augmenté leur nombre, ajouté les dérivées premières et secondes, ajouté une étape de prétraitement (dithering et préaccentuation), ajouté une étape de détection de l'activité vocale et une étape de normalisation par soustraction de cepstre moyen sur fenêtre glissante. Concernant les modèles, nous avons d'abord construit des GMM en utilisant des matrices diagonales de covariance puis nous les avons adaptés en utilisant des matrices pleines de covariances. Nous avons également changé de logiciel d'analyse en choisissant kaldi [Povey et al., 2011] qui est un logiciel spécialement conçu pour la gestion de grandes bases de données parole, c'est le plus utilisé en ce moment dans la reconnaissance de la parole et du locuteur.

Nous avons effectué nos analyses à partir des bases de données complètes présentées partie 4, acquises avec le microphone professionnel, avec le microphone de l'ordinateur et avec le téléphone. Les méthodes que nous avons utilisées et les résultats obtenus ont été présentés à la conférence Interspeech [Jeancolas et al., 2019a] et sont détaillés ci-dessous.

5.1.2 Extraction des MFCC

Comme expliqué dans le paragraphe précédent, nous avons prétraité les enregistrements effectués avec le microphone professionnel et le microphone de l'ordinateur, par soustraction spectrale, afin de supprimer le biais dû au non appariement de l'environnement sonore. Concernant les enregistrements par téléphone, il n'y avait pas de raison que les conditions d'enregistrements

soient différentes entre MP et sains, vu que les sujets appelaient de chez eux avec leur téléphone, nous n'avons par conséquent pas eu besoin d'effectuer l'étape de débruitage par soustraction spectrale. Nous avons extrait la log énergie et 19 MFCC, ainsi que leurs dérivées premières (Deltas) et secondes (Delta-Deltas), sur des fenêtres de 20ms. Ceci conduit à des vecteurs paramétriques de dimension 60 extraits toutes les 10ms. La méthode d'extraction des MFCC est la suivante :

— Prétraitement additionnel : Amélioration de la quantification par *dithering* (ajout d'un bruit stationnaire de manière à supprimer la distorsion due à l'erreur de quantification). Suppression de la composante constante (offset) du signal temporel. Préaccentuation des hautes fréquences avec un coefficient de préaccentuation de 0.97.
— Fenêtrage avec des trames de 20ms et un chevauchement de 50% entre deux trames consécutives. Le fenêtrage utilisé est une fenêtre de type Hamming avec une contrainte de 0 aux extrémités. La fonction fenêtre associée est :

$$f(t) = (0.5 - 0.5 * cos(2\pi * t/T))^{0.85} \qquad avec \quad T = 0.02s \tag{5.1}$$

— Transformée de Fourier rapide de manière à avoir un spectre par trame.
— Banc de filtres de 23 filtres passe bandes MEL triangulaires. L'influence de la plage de fréquence couverte par ces filtres est étudiée dans le paragraphe 5.2.1.5. Le logarithme est appliqué à l'énergie issue des 23 filtres.
— Une transformée en cosinus discrète a enfin été effectuée à partir des log des énergies. Nous avons gardé les 19 premiers coefficients (MFCC), ajouté le log de l'énergie totale de la trame, et calculé les Deltas et Deltas-Deltas, et enfin effectué un liftrage cesptral (mise à l'échelle des coefficients).

Une fois les MFCC et leurs deltas extraits nous avons effectué une détection de l'activité vocale (vad), afin de supprimer les silences. La vad a consisté à garder les trames dont la log énergie respectait la condition suivante :

$$log\ energie > seuil\_vad + m * echelle\_vad \tag{5.2}$$

avec m la moyenne de la log énergie sur l'ensemble du fichier audio, seuil_vad et echelle_vad des constantes valant respectivement 5.5 et 0.5.
Les vecteurs MFCC restants correspondaient ainsi aux trames avec son.

Ensuite une étape de normalisation par soustraction du cepstre moyen (CMS), telle que décrite partie 3.3.1.2 a été réalisée sur des fenêtres glissantes de 300ms centrées autour des trames à normaliser. La moyenne du cepstre est calculée sur chaque fenêtre glissante et soustraite au cepstre de chaque trame. Cette CMS vise à diminuer l'effet convolutionnel linéaire du canal, et diminue ainsi encore plus l'influence de l'inhomogénéité des conditions d'enregistrement.

Pour supprimer un éventuel résidu de biais (dû à un potentiel effet du canal non linéaire) on aurait pu également ajouter une normalisation des MFCC par *Feature mapping* ou une normalisation des scores par HNorm (cf. section 3.3.1.3) si les deux catégories d'environnement acoustique avaient été représentées dans les deux groupes, ce qui n'est pas le cas (l'environnement hors hôpital n'ayant pas concerné les MP). On s'est donc limités à la CMS pour supprimer l'effet du canal.

5.1.3 Entraînement et test des GMM

5.1.3.1 Phase d'entraînement

Tout comme lors de l'analyse préliminaire, nous avons séparé nos sujets en 3 groupes : un groupe avec des sujets MP pour construire le modèle MP, un groupe avec des sujets sains pour

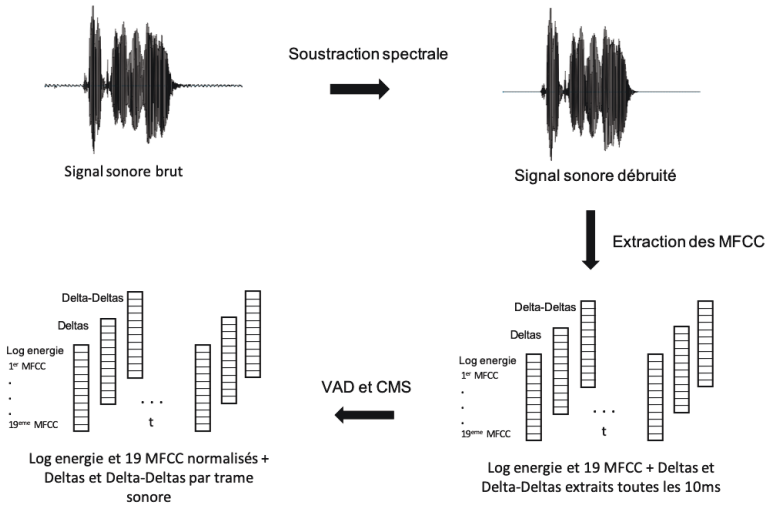


FIGURE 5.5 – Extraction des MFCC

construire le modèle sain et le dernier groupe avec des sujets MP et sains (n'appartenant pas aux groupes d'entraînement), constituant le groupe test. Comme dans l'analyse préliminaire nous avons séparé les hommes des femmes pour l'analyse de manière à optimiser la performance de classification. Deux GMM multidimensionnels ont été construits pour modéliser les distributions des MFCC des groupes MP et sains d'entraînement pour les hommes (après débruitage et VAD). La même chose a été faite en parallèle pour les femmes. Le nombre de gaussiennes choisi a dépendu de la quantité de données utilisées pour l'entraînement. Les GMM ont été construits avec une méthode légèrement différente de la méthode utilisée dans les analyses préliminaires. Nous avons d'abord construit des GMM en utilisant des matrices diagonales de covariance, puis nous avons adapté ces modèles en utilisant des matrices pleines de covariance. Ceci afin d'augmenter la précision sans trop augmenter la puissance nécessaire aux calculs.

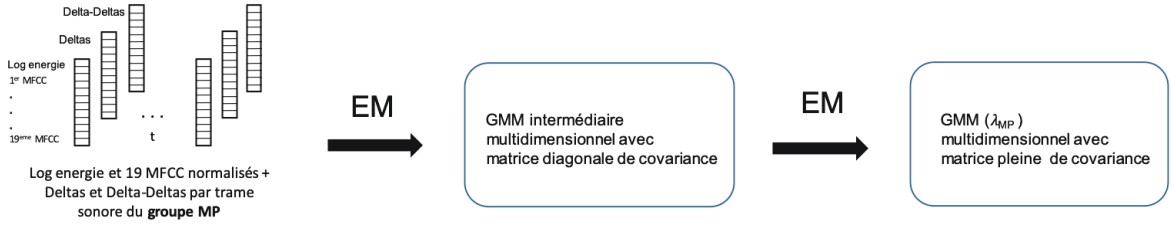


FIGURE 5.6 – Phase d'entraînement du modèle GMM MP. De la même manière un GMM est entrainé à partir des MFCC du groupe sain.

5.1.3.2 Phase de test

Pour chaque sujet test nous avons ensuite calculé pour chaque trame (après VAD) le log de la vraisemblance (LLH) de ses MFCC par rapport aux deux modèles GMM correspondant au genre du sujet. La moyenne sur l'ensemble des trames a ensuite été calculée pour chaque modèle. Une fonction sigmoïde est appliquée sur la différence de ces moyennes (le *log-likelihood ratio*), de manière à produire un score S allant de 0 à 1 par sujet testé, cf. équation 5.3. Un score proche de 1 indique une plus grande probabilité que le sujet testé soit MP et un score proche de 0 qu'il soit sain. On choisira la plupart du temps comme mesure de performance les courbes DET et le taux d'égale erreur, cf. partie 3.1.3.

$$S(X) = \frac{1}{1 + e^{-LLHratio}} \tag{5.3}$$

avec $LLHratio = LLH(X, \lambda_{MP}) - LLH(X, \lambda_{sain})$

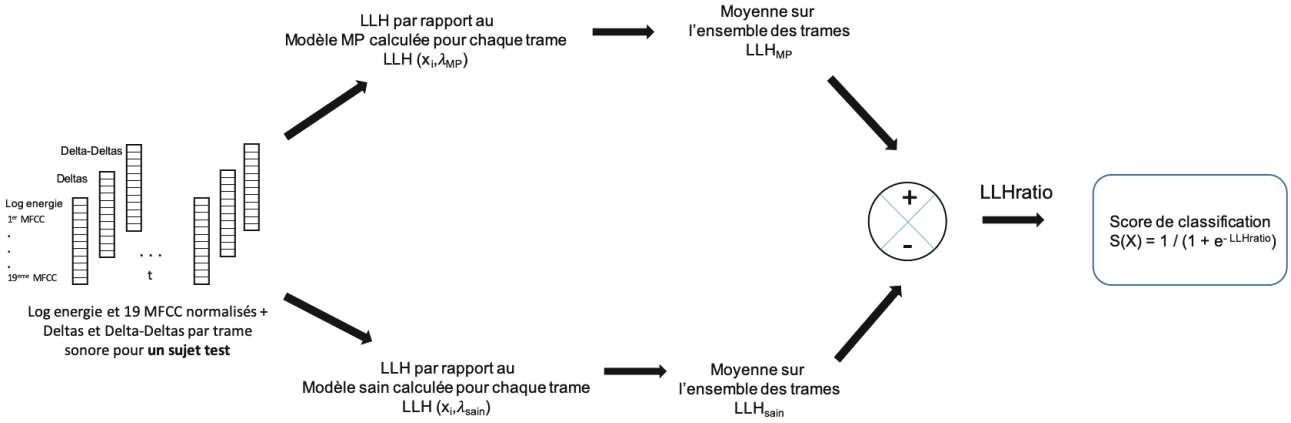


FIGURE 5.7 – Phase de test

5.1.3.3 Phase de validation

Concernant la validation, nous avons choisi une méthode ensembliste de modèles de GMM de type *bootstrap aggregation* sans remise. Cette variante du *bagging* consiste à effectuer une agrégation des scores issus des runs d'un *repeated random subsampling* cf. partie 3.1.5. Nous avons partitionné nos sujets de manière aléatoire en 3 groupes : à savoir deux groupes d'entraînement (un groupe MP et groupe sain) composés chacun d'un nombre fixe et identique de sujets, pour la construction des GMM, et un groupe test constitué des sujets MP et sains restants. Nous avons réitéré ce procédé 40 fois. A l'issu des 40 runs, tous les sujets ont été testés environ une dizaine de fois. Nous avons alors ensuite moyenné par sujet ses scores de classification obtenus pour chaque run où il a été testé, de manière à obtenir un score final Λ par sujet pour toute la base, cf. Figure 5.8.

$$\Lambda_j = \sum_{k=1}^{M} \frac{1}{M} S_{jk}(X) \tag{5.4}$$

M étant le nombre de runs où le sujet j a été dans le groupe test, et S_{jk} son score de classification au k^{ieme} run où il a été testé.

Chaque sujet peut alors être classé par rapport à son score final Λ :

$$\begin{cases} \Lambda_j(X) \geq \theta \text{ alors sujet j est MP} \\ \Lambda_j(X) < \theta \text{ alors sujet j est sain,} \qquad \theta \text{ étant le seuil de décision} \end{cases} \tag{5.5}$$

θ peut être fixé par défaut à 0.5, au taux d'égale erreur, plus haut si on souhaite privilégier la spécificité sur la sensibilité et plus bas si on souhaite privilégier la sensibilité sur la spécificité. Le tracé des courbes DET à partir de ces scores nous a permis d'avoir une vision plus globale de la performance de ce classifieur.

Le choix de cette méthode d'évaluation s'est appuyé sur plusieurs éléments :

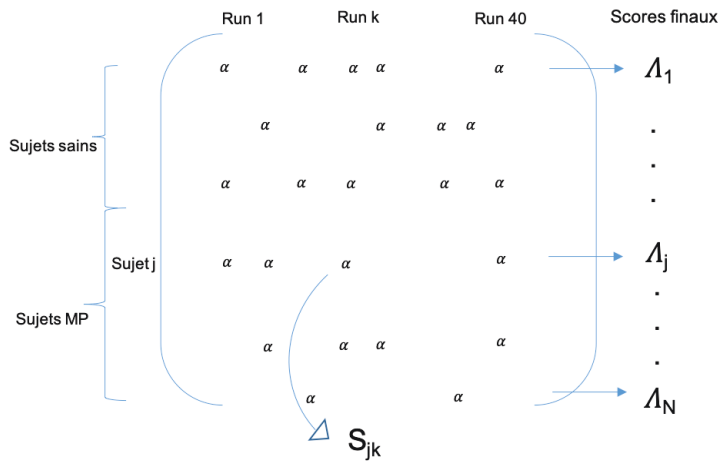


FIGURE 5.8 – Validation de type bootstrap. S_{jk} est le score intermédiaire du sujet j lors du run k (pendant lequel il était dans le groupe test). Λ_j est la moyenne des scores intermédiaires obtenus au cours des 40 runs.

— Tout d'abord, concernant la technique d'échantillonnage, nous avons choisi du repeated random subsampling plutôt que du k-fold car cela nous permettait d'avoir le même nombre de sujets MP et sains pour l'entraînement (ce qui permet d'avoir les mêmes conditions d'entraînement pour les GMM, même nombre optimal de gaussiennes...) et de pouvoir tester à chaque fois tous les autres sujets. Cela n'aurait été possible avec du k-fold ou du LOSO que si nous avions eu autant de MP que de sains dans notre base, ce qui n'est pas le cas.

— Le choix ensuite de compléter cette validation croisée par une méthode ensembliste s'est appuyé entre autres sur le constat d'une variance importante dans les scores de classification d'un même sujet (le score d'un même sujet testé variait en fonction de la composition des groupes d'entraînement). On peut parler de variance "intra-sujet". Les méthodes de type bagging sont connues pour diminuer la variance des prédictions et ainsi améliorer la performance finale de classification [Friedman et al., 2001].

— Concernant le type d'agrégation, nous avons choisi de moyenner les scores plutôt que d'opérer à une agrégation de type vote majoritaire, car c'est la technique qui minimise le plus la variance [Friedman et al., 2001].

— L'utilisation d'une méthode ensembliste nous permettait en plus d'avoir un score final de classification par sujet (ce que ne permettait pas la simple validation croisée *repeated random subsampling*). L'intérêt d'avoir un score par sujet est par la suite de fusionner ce score de classification avec d'autres classifieurs et d'étudier les corrélations avec par exemple des paramètres cliniques et de neuroimagerie.

— Concernant l'erreur réelle (ou généralisée), l'erreur calculée sur les scores finaux (de type *out-of-bag*) est connue pour en être une bonne estimation non biaisée. L'erreur *out-of-bag* estime l'erreur réelle du modèle agrégé de la même manière que l'erreur d'un LOSO estime l'erreur réelle du modèle simple entrainé à partir de toute la base de données. La différence étant que l'erreur réelle du modèle agrégé est souvent meilleure que l'erreur réelle du modèle simple, surtout quand il y a de la variance.

Nous avons choisi 40 pour le nombre de runs car il fallait qu'il soit suffisamment grand pour

que tous les sujets soient testés de préférences plusieurs fois. Nous avons testé des validations avec 80 et 120 runs sans changement apparent des performances. Le nombre de 40 runs s'est donc révélé un bon compromis entre la précision de l'estimation de la performance de notre classifieur et le temps de calcul. Il est courant de trouver des validations avec moins de runs dans la littérature, [Sáenz-Lechón et al., 2006] s'arrêtent à 10 runs par exemple.

Comme mentionné partie 3.1.4, il faut noter que dans tous les cas, que ce soit avec notre approche de type *bagging* ou si on avait choisi une méthode de validation croisée simple, vu que l'on se sert des performances globales de validation pour optimiser certains hyperparamètres (tels que le nombre de gaussiennes, la plage de fréquence utilisée pour l'extraction des MFCC..), nous restons dans le cadre de développement (même si robuste) et non dans le cadre de validation finale. Pour cela il faudrait tester notre modèle final sur une nouvelle base de sujets non vus, pour avoir une estimation précise de la performance réelle de notre classifieur, et utiliser le seuil choisi lors du développement pour la classification finale.

Néanmoins les performances obtenues lors de notre phase de "validation" nous permettent de comparer différents modèles de classification, afin de choisir le plus pertinent, et d'avoir une idée des performances réelles. Nous avons délibérément choisi de ne pas utiliser la méthode ensembliste *boosting* car sans une validation croisée imbriquée ou une base supplémentaire pour estimer l'erreur réelle à la fin, le risque d'*overfitting* aurait été plus grand qu'avec les méthodes de type *bagging*.

5.2 Résultats MFCC-GMM

5.2.1 Résultats avec le microphone professionnel

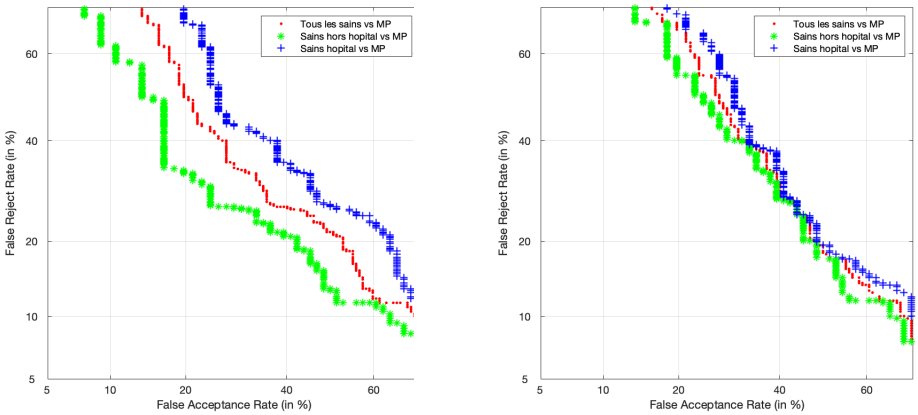
Dans un tout premier temps, afin de valider notre algorithme de classification avec le logiciel kaldi, j'ai effectué une classification par genre. J'ai entrainé un modèle GMM à partir de 36 femmes (équilibré en MP et contrôles) et entrainé un autre GMM à partir de 36 hommes (équilibré en MP et contrôles) et j'ai effectué le test sur les sujets restants. La classification par genre a donné un résultat de 100% de réussite au seuil $\theta = 0.5$.

Pour la classification MP vs sains nous avons analysé les hommes et les femmes séparément, comme expliqué section 5.1.1. Nous commencerons par présenter les résultats obtenus avec les sujets hommes et finirons avec les sujets femmes. Pour la classification des hommes en MP et sain nous avons considéré 36 MP et 36 contrôles pour les groupes d'entraînement et 38 MP et 12 contrôles pour le groupe test. Le choix du nombre de sujets pour chaque groupe a reposé sur la répartition 2/3 pour l'entraînement et 1/3 pour le test concernant les sujets limitants qui sont les sujets sains. Nous avons ensuite réparti les sujets MP de manière à avoir le même nombre que les sujets sains pour les groupes d'entraînement (vu que la précision de la modélisation GMM dépend du nombre de sujets).

5.2.1.1 Comparaison avec et sans débruitage

Afin de tester si le biais du lieu avait été supprimé suite au débruitage (par soustraction spectrale) nous avons tracé les courbes DET avec et sans débruitage, en séparant les sujets sains enregistrés à l'hôpital de ceux enregistrés à l'extérieur. En Figure 5.9 nous avons un exemple avec les résultats pour les tâches de DDK après un run. Sans débruitage on peut constater que la courbe DET des sujets enregistrés à l'extérieur montre de meilleures performances de classification que la courbe des sujets enregistrés à l'hôpital, ce qui traduit le biais identifié auparavant. Après débruitage, on peut voir que les 2 courbes se superposent presque complètement, ce qui signifie que le lieu n'intervient *a priori* plus dans la classification MP vs sain, et donc que le

biais a été supprimé. Ce biais sera retesté à la fin à partir des scores finaux.



(a) Sans débruitage (b) Avec débruitage

FIGURE 5.9 – Courbes DET des sujets tests hommes pour les tâches DDK après un run de classification. Comparaison sans (figure de gauche) et avec débruitage (figure de droite). Les modèles GMM ont été construits à partir de l'ensemble des tâches DDK. Les 11 tâches DDK ont ensuite été testées séparément, donnant 11 scores S(X) par sujet test, soit 550 tests (11*50). Les courbes DET ont été calculées à partir de ces 550 scores.

5.2.1.2 Comparaison des tâches vocales

Premières comparaisons Nous avons d'abord utilisé les MFCC extraits à partir de toutes les tâches (effectuées par le groupe d'entraînement) pour construire les GMM. Cela représentait 3,5 heures de données parole par groupe d'entraînement. Nous avons choisi 500 fonctions gaussiennes pour les GMM. Pour la phase de test nous avons effectué un test par sujet test et par tâche. Les taux d'égale erreur (EER) pour chaque tâche sont énumérés Tableau 5.1. Nous pouvons constater que les tâches avec le plus faible EER (donc la meilleure précision) sont la lecture la plus longue (celle du dialogue), le monologue et les /pataka/. Le fait que le dialogue soit la tâche avec la meilleure performance de la catégorie lecture et répétition de phrases est cohérent avec le fait que c'est la tâche la plus longue de cette catégorie. En effet, en reconnaissance du locuteur, on sait que la qualité des tests dépend entre autres de la quantité de données utilisées pour le test. Le monologue étant d'une durée équivalente à celle du dialogue (environ 1min), et d'un contenu phonétique relativement proche, il est cohérent que son EER soit du même ordre. Concernant les tâches DDK, on peut constater que les deux tâches /pataka/ ont de meilleurs résultats que les autres, ce qui est cohérent avec la littérature. On obtient en moyennant les scores obtenus pour les deux /pataka/ un EER de 22%. La différence entre les deux tâches pataka de 3% illustre la variabilité d'exécution de la même consigne par un même sujet à deux moments différents. Le cas des voyelles soutenues et des tâches de rythme sera discuté par la suite. Le Tableau 5.1 présente aussi les EER pour chaque type de tâche, calculés à partir de la moyenne des scores des sujets obtenus pour chaque tâche de ce type. Nous avons pondéré cette moyenne par la durée des tâches, donnant ainsi plus de poids aux tâches longues, et moins aux tâches courtes. Pour finir nous avons aussi calculé l'EER à partir de la moyenne des scores des sujets issus de chaque tâche, nous avons obtenu $EER = 25\%$. En pondérant la moyenne des scores des sujets par la longueur des tâches nous obtenons une amélioration de 1%, soit $EER = 24\%$. On peut noter que calculer un EER à partir de la moyenne des scores des sujets sur l'ensemble des tâches ne revient pas à moyenner les EER issus de chaque tâche.

voyelles soutenues		DDK		repet lecture		monol		rythme	
/a/	42%	/pa/	34%	lect-phrase1	29%	monol	27%	/pa pa/	31%
/a/ bis	41%	/pou/	38%	lect-phrase2	29%			/kou kou/	37%
glissando	41%	/kou/	36%	lect-texte	29 %			/pa kou/	38%
glissando bis	37%	/poupa/	35%	lect-dialogue	26%			/pa kou/bis	36%
		/pakou/	29%	repet1	38%				
		/pataka/	25%	repet2	29%				
		/pataka/bis	22%	repet3	35%				
		/badaga/	38%	repet4	29%				
		/patikou/	29%						
		/pabikou/	33%						
		/padikou/	39%						
moy : 39%		moy : 27%		moy : 26%		moy : 27%		moy : 32%	
moy tout : 24%									

TABLE 5.1 – EER de la classification des hommes MP vs sain, par tâche vocale, et EER obtenus à partir de la moyenne des scores issus d'un même type de tâches. Les moyennes des scores sont calculées avec une pondération en fonction de la longueur des tâches.

Comparaison GMM généraux et GMM spécifiques Nous voulions ensuite évaluer s'il était possible de mieux adapter le contenu des données d'entraînement utilisées pour les GMM aux tâches utilisées pour le test. Par conséquent, nous avons calculé des GMM plus spécifiques (voir le Tableau 5.2) composé uniquement des tâches que nous avons utilisées pour le test, en modifiant le nombre de fonctions gaussiennes en fonction de la quantité de données de parole. Par exemple, pour savoir si les répétitions de phrases et les tâches de lecture étaient pertinentes pour distinguer les sujets MP des contrôles, au lieu d'utiliser toutes les tâches pour former les GMM, nous avons utilisé uniquement ces dernières. Ce qui fait 54 min de données parole par GMM, pour lesquels le nombre de 20 gaussiennes s'est révélé le plus approprié. Une exception a été faite pour la tâche de monologue, pour laquelle nous avons ajouté des répétitions de phrase et de la lecture au GMM spécifique, afin d'avoir un modèle de parole moins dépendant du contenu. En effet, le contenu du monologue de la cohorte ICEBERG était légèrement différent de celui des sujets témoins externes, de par des différences au niveau du déroulement de leur journée (qu'ils devaient raconter). Nous avons remarqué que les GMM spécifiques amélioraient certaines performances, comme pour les tâches de répétition de phrases et de lecture, ce qui peut s'expliquer par le fait que ces tâches sont texte-dépendant. Au contraire, le monologue, qui est texte-indépendant, n'a pas été améliorée. Nous avons également remarqué que pour d'autres tâches, telles que /pataka/, la performance était diminuée avec des GMM spécifiques (formés avec seulement /pataka/). Cela peut être dû à une quantité réduite de données paroles utilisées pour le GMM (environ 11min de données parole par GMM). Nous avons également testé la tâche /pataka/ avec des GMM formés à partir de toutes les tâches DDK (semi-spécifique) et obtenu les mêmes résultats que pour le GMM général (formé à partir de toutes les tâches). Pour toutes les tâches testées (du moins pour les tâches texte-dépendant), le défi consistait en réalité à trouver le meilleur équilibre entre spécificité et quantité pour les données d'entraînement.

Cas des voyelles soutenues Les tâches de voyelles soutenues, que ce soit avec des GMM spécifiques ou généraux, sont celles qui correspondent à la moins bonne performance (EER=39%). La soustraction du cepstre moyen pouvant altérer la classification quand la variabilité du contenu phonétique est faible (cf. partie 3.3.1.2), nous avons effectué une nouvelle classification sans CMS. Les résultats obtenus ont été similaires. Les performances plus faibles des voyelles soutenues ne sont donc pas la conséquence de la CMS. Ces résultats pourraient être la conséquence de la

faible richesse en contenu phonétique lors de ces tâches. De plus les tâches de voyelles soutenues révèlent surtout des différences liées au timbre, or le timbre est la composante de la voix la plus sensible aux traitements. Sachant que la majorité des sujets MP que nous avons enregistrés étaient sous l'effet de leur traitement, les modifications du timbre, déjà légères au stade débutant, peuvent passer quasiment inaperçues.

5.2.1.3 Fusion

Afin d'améliorer les performances de classification, nous avons effectué une fusion des scores de classification des deux meilleures tâches. Nous avons ainsi moyenné le score de répétition de phrase et de lecture (avec un GMM spécifique) avec le score /pataka/ (avec un GMM global) pour chaque sujet et calculé l'EER. Nous avons obtenu 5% d'amélioration, ce qui a conduit à un EER de 17% ± 5% (voir Tableau 5.2). Comme détaillé partie 3.1.4, l'écart type associé est calculé suivant la formule :

$$\sqrt{EER(1-EER)/n} \tag{5.6}$$

avec n le nombre de sujets sains (facteur limitant).
L'amélioration résultant de la combinaison de ces tâches peut être expliquée par le fait que les types de tâches combinées sont vraiment différents, pouvant ainsi révéler différents types de troubles de la parole liés à la MP (cf. partie 2.3). La combinaison des tâches de répétition de phrases et de lecture avec la tâche de monologue, par exemple, qui sont des tâches relativement proches, n'a pas amélioré les performances.

Tâches testées	Durée test	GMM généraux[1]	GMM spécifiques[2]
Répétitions phrases + lecture	90s	26%	**22%**
Monologue	60s	27%	26%
/pataka/	20s	**22%**	28%
Voyelles soutenues	20s	39%	39%
Toutes les tâches	6min	24%	24%
Fusion[3]		**17%**	

[1] Les GMM généraux ont été entrainés avec toutes les tâches
[2] Les GMM spécifiques ont été entrainés avec les mêmes tâches que celles testées
[3] Fusion des scores des tâches de répétitions de phrases et lecture (testées avec des GMM spécifiques) avec /pataka/ (testée avec des GMM généraux)

TABLE 5.2 – Taux EER pour la classification des hommes MP vs sain à partir des enregistrements de microphone professionnel. Comparaison GMM généraux vs GMM spécifiques pour chaque type de tâches testées.

La matrice de confusion correspondant au seuil de l'EER est détaillée Tableau 5.3 . La matrice de confusion correspondant au seuil *a priori* de 0.5 est présentée Tableau 5.4. L'écart type de la sensibilité et de la spécificité sont définis de la même manière que l'EER avec n égal au nombre de sujets MP pour Se et n égal au nombre de sujets sains pour Sp.

	classés MP	**classés sains**
MP réels	83% ± 4%	17% ± 4%
Sains réels	17% ± 5%	83% ± 5%

TABLE 5.3 – Matrice de confusion de la classification des hommes MP vs sain, au seuil d'EER. Les résultats sont donnés en pourcentage (moyenne ± SD)

	classés MP	**classés sains**
MP réels	$65\% \pm 6\%$	$35\% \pm 6\%$
Sains réels	$8\% \pm 4\%$	$92\% \pm 4\%$

TABLE 5.4 – Matrice de confusion de la classification des hommes MP vs sain, au seuil de 0.5. Les résultats sont donnés en pourcentage (moyenne ± SD)

La Figure 5.10 illustre les distributions normalisées des scores Λ des sujets sains (en bleu) et des sujets MP (en rouge). La courbe DET calculée à partir de ces scores est représentée Figure 5.11. Les barres d'erreur correspondent à $\pm SD$ avec $SD = \sqrt{Err(1-Err)/n}$. Horizontalement *Err=faux positifs* et n=nombre de sujets sains, et verticalement *Err=faux négatifs* et *n=nombre de sujets MP*. Cette courbe est une représentation des performances réelles du modèle agrégé, à partir des scores de fusion.

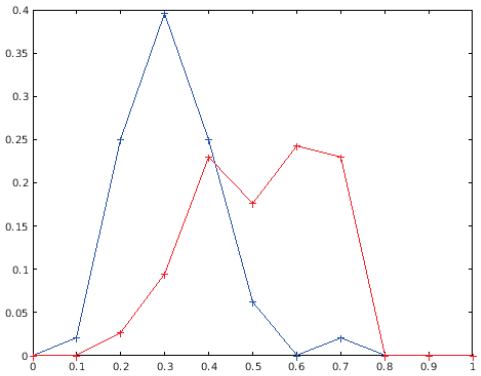


FIGURE 5.10 – Distribution normalisée des scores Λ de fusion, en bleu (courbe de gauche) pour les sujets hommes sains et en rouge (courbe de droite) pour les sujets hommes MP.

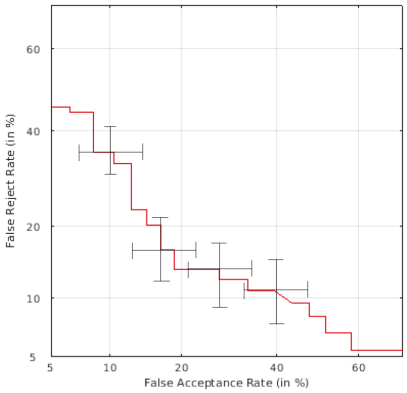


FIGURE 5.11 – Courbe DET à partir des scores Λ de fusion issus de la classification des hommes MP vs sain. Les barres d'erreurs ont été calculées à partir des écarts types.

5.2.1.4 Influence du nombre de sujets pour l'entraînement des GMM

Les performances diminuent quand on prend moins de 20 sujets par groupe d'entraînement. Entre 30 et 40 sujets par groupe d'entraînement, on n'observe pas beaucoup de différence dans les performances. On ne peut pas prendre beaucoup plus que 40 sujets car on est limité par

notre nombre de sujets sains, et on veut le même nombre de sujets pour l'entraînement MP et l'entraînement sain. D'autre part plus on prend de sujets pour l'entraînement, moins il en reste pour les tests et donc si on veut suffisamment de tests par sujet (pour diminuer la variance) il faut augmenter le nombre de runs, ce qui signifie augmenter le temps de calcul.

5.2.1.5 Influence des plages de fréquences, des fréquences d'échantillonnage et du nombre de filtres

Dans un premier temps nous avions testé de prendre pour la plage de fréquence des MFCC celle de notre microphone professionnel, soit [50-20000Hz]. En prenant 23 filtres MEL, qui est le nombre de filtres le plus courant et en gardant la fréquence d'échantillonnage du microphone (96kHz). Nous avons obtenu un EER de 30% à partir des scores de fusion. Nous avons testé d'augmenter le nombre de filtres, car le nombre de 23 filtres est surtout utilisé pour les enregistrements de 16kHz, mais ceci n'a pas entrainé d'amélioration des performances. Nous avons ensuite étudié l'influence de la bande de fréquence. Les meilleurs résultats ont été obtenus pour la bande [20-7000Hz], avec un EER de 17%, soit une amélioration de 13%. Tous les résultats présentés dans les paragraphes précédents ont été obtenus à partir de cette plage de fréquence. Abaisser la limite inférieure à 20Hz permet ainsi de prendre en compte les fréquences enregistrées les plus basses (comprises entre 50 et 70Hz), qui ne seraient pas bien prises en compte si le début du premier filtre triangulaire commence à 50Hz. Diminuer la limite supérieure permet de donner plus d'importance aux fréquences inférieures à 7000Hz qui contiennent la partie la plus importante des fréquences vocales. Cependant si on restreint encore plus la bande de fréquence pour le calcul des MFCC, en prenant par exemple la bande de fréquence utilisée avec les téléphones, soit [300-3700Hz], on peut constater une dégradation des performances de 8%.

La fréquence d'échantillonnage, pour une même bande de fréquence semble avoir peu d'influence dans les résultats de classification. On n'observe pas de changement significatif dans les résultats entre 96kHz et 16kHz, et une détérioration de 1 à 2 % quand on passe à 8kHz. Néanmoins la bande de fréquence [20-7000Hz] ne peut pas être testée avec une fréquence d'échantillonnage de 8kHz, car le théorème de Shannon impose que la limite supérieure soit inférieure à la moitié de la fréquence d'échantillonnage. Donc une fréquence d'échantillonnage de 8kHz impose d'avoir une limite supérieure d'environ 3700Hz pour le calcul des MFCC, et on a vu que cette limite donnait de moins bons résultats que la limite de 7000Hz.

5.2.1.6 Comparaison du modèle agrégé avec le modèle simple

Comme expliqué à la section 3.1.5 en théorie le modèle agrégé devrait avoir de meilleures performances de classification que le modèle simple. Pour vérifier cela, nous avons tracé les distributions des scores issus des deux modèles, calculé leur courbes DET et leur EER. Pour la comparaison nous avons effectué les classifications à partir des tâches de lecture et répétitions de phrase testées avec des GMM spécifiques (entrainés à partir de ces mêmes tâches). Les barres d'erreur des courbes DET sont calculées à partir des écarts types.

Nous avons représenté Figure 5.12 la distribution des scores α de classification des sujets tests pour chaque run, ainsi qu'en prenant les scores de tous les runs cumulés. En bleu sont représentés les sujets sains réels et en rouge les sujets MP réels. Cette figure donne une visualisation des scores correspondant au modèle simple. Enfin la Figure 5.13 illustre la distribution des scores Λ (calculés en moyennant les scores α des sujets issus de chaque run où ils ont été testés), ces scores sont ceux du modèle agrégé (cf. section 5.1.3.3).

La Figure 5.14 représente la courbe DET de test du modèle agrégé (à partir des scores Λ). Comme vu partie 3.1.5, cette courbe est une bonne estimation des performances réelles du

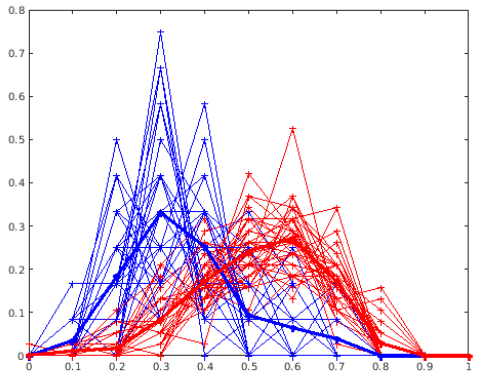


FIGURE 5.12 – Distribution normalisée des scores α de classification des hommes MP vs sain, à partir des tâches de lecture et répétitions de phrases enregistrées avec le microphone professionnel. Les distributions de chaque run sont représentées en traits fins et les distributions de l'ensemble des scores α, tous runs confondus, sont représentées en traits épais. En bleu sont représentés les sujets sains réels et en rouge les sujets MP réels.

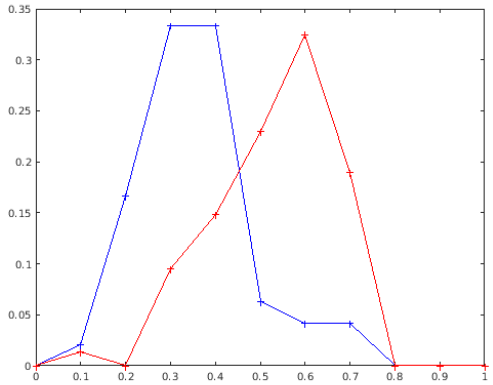


FIGURE 5.13 – Distribution normalisée des scores Λ de classification des hommes MP vs sain, à partir des tâches de lecture et répétitions de phrases. En bleu sont représentés les sujets sains réels et en rouge les sujets MP réels.

modèle agrégé. L'EER obtenu est de $22\% \pm 6\%$. L'écart type étant calculé à partir de la formule 5.2.1.3. Nous avons également calculé l'EER d'entraînement du modèle agrégé, en testant pour chaque run les sujets du groupe d'entraînement. L'EER obtenu à partir des sujets d'entraînement est 0%. Ceci montre l'importance d'avoir des sujets dans le groupe test n'ayant pas servi pour l'entraînement afin d'avoir une estimation non biaisée de l'erreur réelle.

Les courbes DET calculées à partir des scores α des sujets tests pour chaque run sont tracées Figure 5.15. La moyenne de ces courbes DET est également tracée (à partir de la moyenne des taux FP et FN obtenus pour chaque run, ce pour chaque incrémentation du seuil). La courbe DET moyenne représente, comme expliqué dans [Sáenz-Lechón et al., 2006], les performances réelles du modèle simple. Dans notre configuration, vu que les nombres de MP et de sains testés sont identiques d'un run à l'autre, cela revient exactement à calculer directement la courbe DET à partir de l'ensemble des scores α (tous les runs cumulés).

L'estimation de l'EER réel, correspondant à la moyenne des EER de chaque run est de $23.6\% \pm 6\%$. L'écart type associé à cette estimation est la valeur de l'écart type des EER de

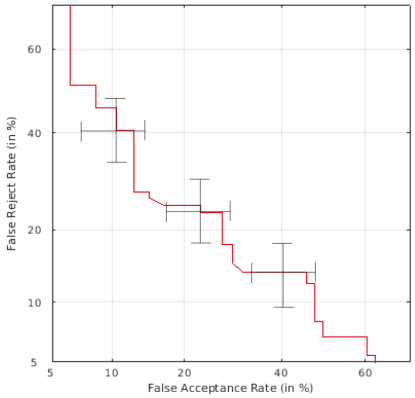


FIGURE 5.14 – Courbe DET à partir des scores Λ issus de la classification des hommes MP vs sain, à partir des tâches de lecture et répétitions de phrases. Les barres d'erreurs ont été calculées à partir des écarts types.

chaque run. L'EER calculé à partir de la DET moyenne (ou à partir de l'ensemble des scores α) est de 24.1%±6%. Même si l'estimation de l'EER en prenant la moyenne des EER est légèrement plus optimiste que celle effectuée sur la DET moyenne, elles restent très proches.

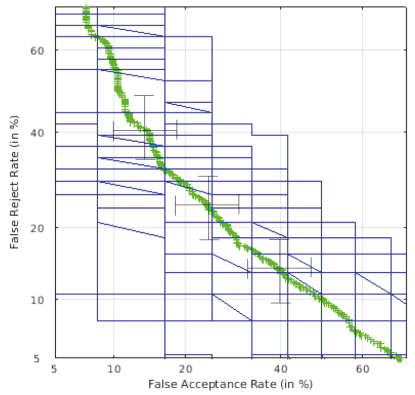


FIGURE 5.15 – Courbes DET issues de la classification des hommes MP vs sain, à partir des tâches de lecture et répétitions de phrases enregistrées avec le microphone professionnel. Les courbes en bleu sont calculées à partir des scores α des sujets tests pour chaque run. La courbe verte épaisse correspond à la moyenne. Les barres d'erreurs ont été calculées à partir des écarts types.

Enfin Figure 5.16 nous avons tracé les courbes DET correspondant au modèle agrégé et au modèle simple sur un même graphe pour faciliter la comparaison. Les écarts entre les courbes DET et la diminution de l'EER de 2% du modèle agrégé par rapport au modèle simple confirme bien la légère amélioration attendue résultant de l'agrégation des modèles de chaque run [Friedman et al., 2001].

5.2.1.7 Effet des traitements

Tous les patients MP traités ayant été enregistrés sous l'effet de leur traitement, à savoir moins de 12 heures après leur prise médicamenteuse du matin, il est compliqué d'évaluer l'effet du traitement médicamenteux sur les performances de classification, d'autant que tous les trai-

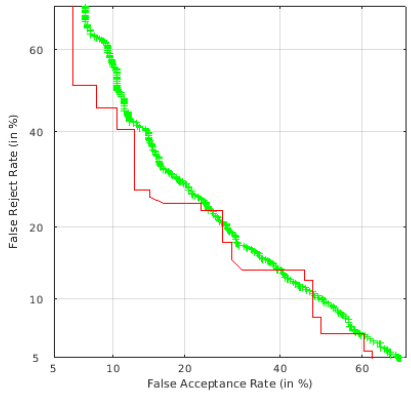


FIGURE 5.16 – Comparaison des courbes DET représentant les performances du modèle agrégé et du modèle simple, lors de la classification des hommes MP vs sain, à partir des tâches de lecture et répétitions de phrases enregistrées avec le microphone professionnel. Le modèle agrégé est représenté par la courbe DET rouge calculée à partir des scores Λ. Le modèle simple est représenté par la courbe DET verte calculée à partir de la moyenne des DET issues de chaque run.

tements n'ont pas la même durée d'action et que tous les patients n'assimilent pas à la même vitesse leurs traitements. Concernant les thérapies vocales 13 patients ont reporté avoir suivi une thérapie de type LSVT au cours des deux dernières années. Nous avons calculé les performances de classification de ces patients et avons obtenu un EER de 15% en prenant en compte le score de fusion. Cette diminution de l'EER en ciblant les patients suivant une thérapie de type LSVT semblerait indiquer à première vu que l'effet de la thérapie vocale est négatif. En fait cela indique surtout que les patients qui ont choisi de suivre une thérapie LSVT sont ceux qui sont les plus gênés par leur voix et donc, *a priori* ceux qui ont le plus de problèmes vocaux. Pour évaluer réellement l'effet de la thérapie vocale il aurait fallu imposer à un groupe de patients sélectionnés au hasard de faire cette thérapie, ou bien enregistrer les patients avant puis après leur thérapie et analyser la différence.

5.2.1.8 Effet du lieu d'enregistrement

Comme expliqué partie 5.1.1, nous avons effectué une soustraction spectrale avant d'extraire les MFCC pour limiter l'impact du lieu d'enregistrement sur la classification. Afin de vérifier que le lieu d'enregistrement n'a plus d'effet significatif sur les classifications, nous avons comparé les performances des sujets enregistrés à l'hôpital à celles des sujets enregistrés hors hôpital. Les différences avec les performances calculées sur tous les sujets n'excèdent pas 3%, et les t-tests que nous avons effectués n'ont pas montré d'écart significatif entre les différents lieux d'enregistrements.

5.2.1.9 Classification des femmes

Pour les femmes, nous avons considéré 30 sujets MP et 30 sujets sains pour l'entraînement et 11 sujets MP et 13 sujets sains pour le test. Nous avons utilisé les mêmes hyperparamètres que chez les hommes pour entraîner les modèles, à l'exception du nombre de gaussiennes que nous avons diminué, pour correspondre à la quantité de données plus réduite servant à entraîner les modèles. Nous avons obtenu un taux d'EER de $42\% \pm 8\%$ pour les tâches de répétition de phrase et de lecture (avec un GMM spécifique), un taux d'EER de $47\% \pm 8\%$ pour la tâche /pataka/ (avec DDK utilisé pour les GMM) et un taux d'EER de $45\% \pm 8\%$ pour le monologue. Ces résultats sont nettement moins bons que les équivalents obtenus avec les hommes, et nous

avons observé une plus grande variabilité entre les différents runs.

Afin d'être sûrs que le problème ne vient pas de la quantité plus réduite de données disponibles, nous avons effectué une classification de contrôle avec les sujets hommes en utilisant le même nombre de sujets que pour les femmes pour l'entraînement et le test. Les résultats obtenus étaient sensiblement équivalents à ceux présentés dans la partie précédente. La quantité plus réduite de données disponible pour les femmes n'explique donc pas la mauvaise performance. Nous avons également testé les sujets femmes par rapport aux modèles entrainés avec les modèles hommes, ce qui n'a pas non plus amélioré les performances. Finalement nous avons aussi testé la classification des femmes en considérant une longueur de trame plus courte (15ms au lieu de 20ms) car il a été montré dans certaines études que des trames temporelles plus courtes pour les femmes conduisaient à de meilleures performances dans la reconnaissance du locuteur [Li and Zheng, 2015], ce qui n'a pas entrainé d'améliorations.

Ainsi nous pouvons en déduire que notre algorithme de classification n'est pas adapté pour les femmes, dont on sait que les MFCC connaissent plus de variabilité que pour les hommes [Fraile et al., 2009b]. Les auteurs de [Tsanas et al., 2011] avaient aussi trouvé que les MFCC semblaient plus adaptés pour suivre l'évolution de MP chez les hommes que chez les femmes, suggérant un processus de dégradation de la voix dans MP différent entre les hommes et les femmes.

Notre moins bonne détection de MP chez les femmes peut aussi être expliquée en partie par une atteinte neuronale en moyenne moins marquée chez les femmes MP que chez les hommes MP [Haaxma et al., 2007] et une symptomatologie plus bénigne chez les femmes, d'après cette même étude. De plus, d'après la littérature, une apparition des symptômes de MP plus tardive de 2 ans en moyenne serait observée chez les femmes, comparé aux hommes. L'effet éventuellement protecteur de l'œstrogène sur la maladie de Parkinson a souvent été avancé pour expliquer les différences dans l'expression de la MP selon le genre.

Nous pouvons d'ailleurs constater dans notre base de données que le score moteur de l'UPDRS III des sujets MP est en moyenne plus élevé chez les hommes que chez les femmes (on a un score moyen de 34 pour les hommes et 29 pour les femmes). Les hommes et les femmes étant appariés en âge dans notre base de données, cette différence est cohérente avec le développement plus tardif des symptômes moteurs chez les femmes, et peut contribuer à expliquer la plus grande difficulté à détecter MP chez les femmes dans notre base de données.

Enfin une autre cause possible, pouvant contribuer à expliquer la moins bonne performance de détection de MP chez les femmes, pourrait venir des supports neuronaux de la parole. En effet des études ont montré que les circuits neuronaux de la parole étaient différents chez les hommes et chez les femmes [de Lima Xavier et al., 2019,Jung et al., 2019]. Ces circuits neuronaux peuvent donc être différemment impactés par la MP, et conduire à différents types ou différents degrés d'altérations vocales suivant le genre.

5.2.2 Résultats avec le microphone de l'ordinateur

5.2.2.1 Classification des hommes MP vs sain

Pour nos enregistrements effectués avec le microphone de l'ordinateur nous avons pris les mêmes nombres de sujets que pour le microphone professionnel (vu que les mêmes sujets ont été enregistrés avec les 2 microphones). Soit 36 MP hommes et 36 contrôles hommes pour les groupes d'entraînement et 38 MP hommes et 12 contrôles hommes pour le groupe test.

L'intérêt des enregistrements avec le microphone de l'ordinateur est qu'il permet d'évaluer la pertinence de la qualité du microphone. Les enregistrements avec le microphone professionnel

et le microphone de l'ordinateur ont été effectués simultanément. La base de données est donc strictement identique entre les deux microphones. Nous avons néanmoins constaté que les enregistrements issus du microphone de l'ordinateur avaient tendance à se déclencher entre 0 à 1s plus tard, résultant de la gestion automatisée en série des enregistrements par notre interface GUI. Après vérification, cela n'a entrainé la perte de quasiment aucune donnée vocale.

Les principales différences entre les enregistrements du microphone professionnel et ceux du microphone interne de l'ordinateur (outre les différences de qualités de microphones) résident dans la présence d'une fonction de réduction de bruit de fond activée par défaut pour les ordinateurs de type Macbook, et la distance entre le microphone et la bouche (environ 50cm au lieu de 5 à 10cm pour le microphone professionnel).

Les EER obtenus avec le microphone intégré de l'ordinateur sont présentés Tableau 5.5. Nous pouvons constater une dégradation de 4% concernant la performance à partir des tâches de lecture et répétition de phrase et une dégradation de 12% concernant la tâche /pataka/. Les caractéristiques vocales associées à l'exécution de cette tâche très spécifique par la population MP semblent donc plus affectées par la qualité du microphone. De même le score de fusion pâtit ici de la mauvaise performance associée à la tâche /pataka/. La dégradation des performances avec le microphone interne de l'ordinateur peut venir du microphone de moins bonne qualité ainsi que de la fonction de débruitage actif de l'ordinateur ou de la distance accrue entre la bouche et le microphone.

	lecture + repet	/pataka/	fusion
micro professionnel	22%	22%	17%
micro ordinateur	26%	35%	32 %

TABLE 5.5 – Comparaison de l'EER issu de la classification des hommes MP vs sain avec le microphone professionnel et avec le microphone de l'ordinateur. Les tâches de lecture et répétition de phrases ont été testées à partir de GMM spécifiques, et la tâche /pataka/ à partir de GMM généraux. Pour la fusion le taux est obtenu en moyennant les deux scores de classification (lecture + répétition de phrases et /pataka/) pour chaque sujet.

5.2.2.2 Effet "cross-micro"

Nous avons également voulu évaluer les performances lors d'une classification *"cross micro"*, c'est à dire estimer la performance de classification de sujets enregistrés avec un microphone différent de celui utilisé pour entraîner les GMM. Pour cela nous avons entrainé les GMM avec les enregistrements issus du microphone professionnel et nous avons utilisé les enregistrements du microphone de l'ordinateur pour les sujets tests. Nous avons obtenu un EER de 25% pour les tâches de lecture et répétition de phrases et 37% pour la tâche /pataka/, soit des résultats semblables à ceux obtenus en entrainant les GMM avec le microphone de l'ordinateur.

Pour conclure, lorsque les conditions d'enregistrement sont telles que la distance avec le microphone est grande et/ou qu'une fonction de réduction de bruit de fond est active il semble préférable d'utiliser seulement les tâches de lecture et répétitions de phrases. En procédant ainsi le classifieur MFCC-GMM conduit à un EER d'environ 25%.

5.2.2.3 Classification des femmes MP vs sain

Pour la classification des femmes, nous avons suivi la même méthode qu'avec le microphone professionnel, et avons obtenu un EER de $42 \pm 8\%$ pour les tâches de lecture et répétition

de phrase et de $45 \pm 5\%$ pour la tâche /pataka/. Ces taux d'erreur sont comparables à ceux obtenus avec le microphone professionnel. Nous n'observons pas de dégradation significative contrairement aux hommes, ce qui n'est pas surprenant vu que les performances étaient déjà très faibles, de plus les marges d'erreur des EER sont plus grandes que pour les hommes.

5.2.3 Résultats avec le téléphone

Tout comme pour les résultats obtenus avec le microphone professionnel, et le microphone de l'ordinateur, nous nous concentrons principalement sur les analyses effectuées avec les hommes. La raison étant que cette méthode d'analyse est peu appropriée à la détection de MP chez les femmes, dont la plus grande variabilité des MFCC semble nuire à cette méthode de classification.

Pour rappel, 63 MP hommes et 36 sujets sains hommes ont participé aux enregistrements téléphoniques mensuels, faisant un total de 1 à 13 sessions par sujet, avec une moyenne de 5 sessions par participant. Les tâches vocales effectuées sont de même type mais un peu moins nombreuses que les tâches vocales effectuées à l'hôpital. Elles ne contenaient pas de lecture, car pour des raisons pratiques nous voulions que toutes les consignes soient audio. Quatre tâches DDK ont été enlevées pour les enregistrements téléphoniques et la tâche /pataka/ n'est effectuée qu'une seule fois (au lieu de deux). Ceci dans le but que les sessions téléphoniques ne soient pas trop longues, afin de préserver la motivation des participants pendant toute la durée du protocole.

La fréquence d'échantillonnage des enregistrements était de 8kHz et la bande de fréquences considérée : [300-3700]Hz. Concernant les analyses, les MFCC ont été extraits de la même manière qu'avec les enregistrements effectués à l'hôpital, à l'exception de la soustraction spectrale qui était moins justifiée. En effet les conditions d'enregistrement n'étaient pas différentes entre les groupes MP et sain, et contrairement aux enregistrements effectués à l'hôpital, nous n'avions pas de tâche de silence à utiliser pour la soustraction spectrale. La soustraction du cepstre moyen a quant à elle été préservée, ce qui limite les effets des canaux.

Pour l'entraînement des GMM nous avons considéré des groupes de 30 MP et 30 sains, les autres sujets étant utilisés pour le test (33 MP et 6 contrôles).

5.2.3.1 Comparaison enregistrements micro pro vs simulation téléphone

Dans un premier temps considérons une simulation grossière du téléphone, en sous échantillonant les enregistrements issus du microphone professionnel (8kHz au lieu de 96kHz) et en réduisant la bande de fréquence utilisée pour le calcul des MFCC ([300-3700]Hz au lieu de [20-7000]Hz). Le but étant d'analyser l'effet de ces deux caractéristiques propres au téléphone qui sont connues pour dégrader le signal et les analyses basées sur les MFCC [Wu et al., 2018, Fraile et al., 2009a]. On observe une dégradation de 9% de l'EER en prenant les scores de fusion (cf. Tableau 5.6 et Figure 5.17).

5.2.3.2 Comparaison simulation téléphone vs enregistrements téléphoniques réels

En ce qui concerne les enregistrements téléphoniques réels, pour comparer leurs résultats avec les résultats précédents nous avons considéré comme résultat de classification finale le score de DDK et non le score de fusion. En effet la fusion telle qu'elle a été réalisée pour les enregistrements professionnels (moyenne des scores obtenus avec répétition de phrase + lecture et tâche /pataka/) n'était pas pertinente pour les enregistrements téléphoniques en raison de l'absence de lecture et des tâches DDK moins bien adaptées aux tests des /pataka/.

Lorsque toutes les sessions téléphoniques ont été utilisées pour les groupes d'entraînement et qu'une seule session a été utilisée par sujet test, on constate une perte de performance de 9% par rapport à la simulation téléphonique. Cela pourrait s'expliquer par une augmentation du bruit de fond pour le téléphone et des distorsions d'amplitude qui sont aussi connus pour détériorer les analyses basées sur les MFCC [Vásquez-Correa et al., 2017b, Fraile et al., 2009a]. La diminution des performances peut également trouver une explication dans la qualité d'exécution des tâches. Aucun expérimentateur n'était présent lors des enregistrements téléphoniques pour faire recommencer les tâches lorsque les instructions n'étaient pas respectées. En ce qui concerne le nombre légèrement inférieur de sujets dans la base téléphonique, cela devrait être compensé par le nombre plus important de sessions utilisées par sujet pour l'entraînement des GMM. Il faut néanmoins étudier la comparaison entre les enregistrements téléphoniques réels et les enregistrements téléphoniques simulés avec précaution, vu que les tâches analysées diffèrent légèrement.

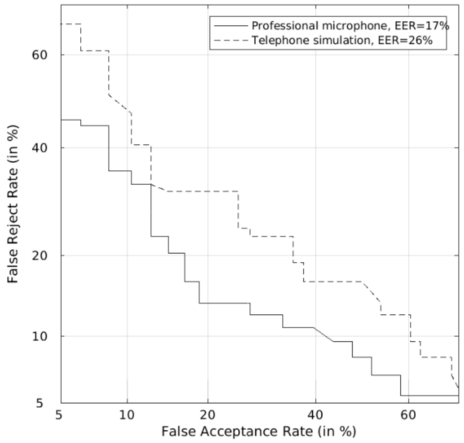


FIGURE 5.17 – Courbes DET issue de la classification hommes MP vs hommes sains à partir des scores de fusion. Comparaison entre microphone professionnel et simulation du téléphone.

5.2.3.3 Influence du nombre sessions pour les enregistrements téléphoniques réels

Afin d'évaluer l'influence de la quantité de données de test sur les performances des enregistrements téléphoniques, nous avons effectué une classification en utilisant cette fois toutes les sessions téléphoniques réalisées par sujet de test et nous avons comparé les résultats à la classification effectuée en utilisant une seule session par sujet de test. Nous avons constaté des améliorations de la classification, atteignant un **EER de** $25\% \pm 7\%$ **pour les tâches DDK** (voir Tableau 5.6 et Figure 5.18). Pour les tâches DDK, les résultats obtenus étaient encore meilleurs que ceux obtenus avec le microphone professionnel. Cela signifierait qu'avec environ 5 minutes de données paroles téléphoniques DDK par personne, contre environ 1 minute 30 pour le microphone professionnel, l'amélioration de la performance due à l'augmentation de la quantité a prévalu sur la dégradation de la qualité.

Avec nos enregistrements issus du microphone professionnel, tester les tâches /pataka/ avec des GMM globaux semble plus pertinent que d'utiliser l'ensemble des tâches DDK pour le test (en prenant la moyenne de scores). Curieusement, avec nos vrais enregistrements téléphoniques, cela semble être le contraire. Cela pourrait s'expliquer par le contenu des tâches DDK téléphoniques moins adaptées au test de la tâche /pataka/ que les données DDK de nos enregistrements professionnels.

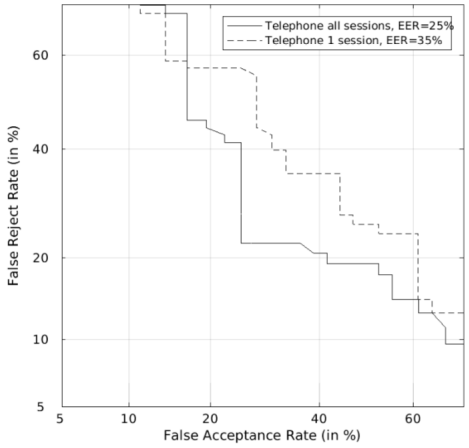


FIGURE 5.18 – Courbes DET issue de la classification hommes MP vs hommes sains à partir des scores de DDK issus des vrais enregistrements téléphoniques. Impact de la quantité de données pour le test : comparaison entre une seule session testée par sujet et toutes les sessions testées par sujet.

Tâches testées[1]	micro professionnel	Téléphone simulation	Téléphone 1 sessions	Téléphone toutes sessions
Répétitions phrases[2]	22%	28%	37%	36%
Monologue	26%	29%	39%	36%
DDK	31%	32%	35%	25%
/pataka/	22%	25%	42%	32%
Résultats finaux[3]	**17%**	**26%**	**35%**	**25%**

[1] La répétition de phrases, le monologue et DDK sont testées avec des GMM spécifiques, et /pataka/ est testée avec des GMM généraux
[2] Pour les enregistrements issus du microphone professionnel et de la simulation du téléphone, les tâches de lecture sont ajoutées aux répétitions de phrases pour l'entraînement et le test
[3] Les résultats finaux sont composés du résultat de fusion pour le microphone professionnel et la simulation du téléphone, et du résultat de DDK pour les enregistrements téléphoniques réels

TABLE 5.6 – Taux EER pour la classification MP hommes vs sains hommes. Comparaison entre les enregistrements du microphone professionnel, la simulation du téléphone, et les enregistrements téléphoniques réels (en prenant toutes les sessions pour l'entraînement des GMM et soit une, soit toutes les sessions pour les sujets tests).

Nous avons voulu étudier de plus près l'influence du nombre de sessions d'enregistrements téléphoniques effectuées par les sujets sur la performance de leur classification. Le Tableau 5.7 détaille le taux de sujets bien classés en fonction du nombre de sessions téléphoniques qu'ils ont effectuées. Nous pouvons observer une amélioration importante dans les performances de classification à partir de 5 sessions par sujet. En effet si on considère seulement les sujets ayant effectué 4 sessions ou moins, l'EER issu des tâches DDK est de 44%, alors que si on considère les sujets ayant effectué 5 sessions ou plus, l'EER est de 19%. Cette différence très importante est à relativiser car les performances ont été calculées à partir de sujets différents, et le nombre faible de sujets ayant effectué 4 sessions ou moins (23 sujets) donne une valeur peu précise de cet EER. Pour étudier l'influence du nombre de sessions sur les performances de manière pertinente et non biaisée, nous avons alors considéré seulement les sujets ayant effectué 5 sessions ou plus. Nous avons calculé l' EER de ce groupe lorsque toutes leurs sessions étaient prises en compte, et obtenu 19% comme décrit précédemment, et nous avons calculé leur EER lorsque une seule session de ces mêmes sujets était prise en compte, et avons alors obtenu 28%. Nous constatons alors une différence de 9% suivant qu'une session téléphonique est utilisée par sujet ou que 5

sessions ou plus sont utilisées. Nous retrouvons l'ordre de grandeur de l'amélioration trouvée précédemment lorsque toutes les sessions (allant de 1 à 13) des sujets testés étaient utilisées versus une seule session était utilisée par sujet testé.

nb sessions	1	2	3	4	5	6	7	8	9	10	11	12	13
$\frac{nb\ sujets\ bien\ classes}{nb\ sujets\ testes}$	$\frac{3}{8}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{3}{5}$	$\frac{6}{6}$	$\frac{2}{3}$	$\frac{6}{6}$	$\frac{5}{8}$	$\frac{8}{8}$	$\frac{12}{14}$	$\frac{12}{16}$	$\frac{6}{8}$	$\frac{5}{7}$

TABLE 5.7 – Taux de bonnes classifications des sujets en fonction du nombre de sessions téléphoniques effectuées.

5.2.3.4 Influence du téléphone portable vs téléphone fixe

Les participants de notre base de données téléphoniques ont effectué les enregistrements à partir de leur téléphone personnel, sans contrainte sur le fait d'utiliser un téléphone fixe ou un téléphone portable. Parmi les participants hommes 33 sujets sains ont opté pour leur téléphone portable, et 3 pour un téléphone fixe. De même 54 sujets MP ont opté pour leur téléphone portable, et 9 pour leur téléphone fixe. Si on considère les résultats de classification MP homme vs sain homme pour la tâche DDK, nous avions obtenu un EER de 25% en prenant tous les sujets. Si maintenant on considère seulement les sujets ayant utilisé un téléphone fixe, nous obtenons pour le même seuil un taux de faux positifs et de faux négatifs de 33%. Cette légère dégradation est à considérer avec précaution car le nombre de personnes ayant utilisé leur téléphone fixe est très faible. Néanmoins si cette détérioration est réelle, elle pourrait éventuellement provenir du fait que l'entraînement des GMM à été effectué en grande majorité avec des enregistrements de téléphones portables, ce qui pourrait conduire à un pouvoir discriminant plus faible pour détecter MP à partir d'enregistrements issus de téléphones fixes.

5.2.3.5 Comparaison du modèle agrégé vs modèle simple

Pour finir tout comme pour les enregistrements issus du microphone professionnel, nous avons estimé la performance du modèle simple, pour la comparer à la performance du modèle agrégé. Pour cela nous avons utilisé les tâches DDK de toutes les sessions téléphoniques, pour lesquelles nous avions obtenu un EER=25% avec le modèle agrégé. Nous avons calculé l'EER du modèle simple, en suivant la méthode décrite partie 5.2.1.6 et avons obtenu un EER de 28%. On retrouve ainsi une légère amélioration des performances résultant du modèle agrégé.

5.2.3.6 Classification des femmes

Concernant les femmes, nous avons analysé les données téléphoniques de 38 sujet MP et 25 sujets sains. Pour chaque run de classification nous avons considéré 20 sujets MP et 20 sains pour l'entraînement des modèles GMM, et le reste des sujets, soit 18 MP et 5 sains pour le test.

Nous avons obtenu un EER de $41 \pm 10\%$ pour les tâches DDK ainsi que pour le monologue. Ces taux d'erreur sont comparables à ceux obtenus avec le microphone professionnel. Nous n'observons pas de dégradation significative contrairement aux hommes, ce qui n'est pas surprenant vu que les performances étaient déjà très faibles, de plus les marges d'erreur des EER sont plus grandes que pour les hommes.

5.2.4 Classification des iRBD

Nous avons dans un premier temps effectué une classification iRBD vs sain pour les hommes à partir des enregistrements issus du microphone professionnel. De la même manière que pour la

classification MP vs sain nous avons entrainé pour chaque run un modèle GMM iRBD (composé de 30 sujets iRBD) et un modèle GMM sain (composé de 30 sujets sains) et nous avons calculé la vraisemblance des sujets restants (11iRBD et 18 sains) par rapport à ces deux modèles. Les EER à l'issue des 40 runs sont donnés Tableau 5.8 pour les tâches de lecture et répétition de phrase, pour la tâche /pataka/ et pour la fusion des deux. Nous pouvons constater une diminution moyenne des performances d'environ 20% par rapport à la classification MP vs sain. Cette différence est cohérente avec le fait que les iRBD, considérés comme des pré-Parkinsoniens, présentent moins de troubles vocaux que les Parkinsoniens avérés.

Nous avons dans un deuxième temps effectué la même classification iRBD vs sain en utilisant cette fois les enregistrements issus de microphone intégré de l'ordinateur. Nous avons obtenu un EER de 45%, soit une dégradation de 8% par rapport aux enregistrements issus du microphone professionnel.

	lecture + repet	/pataka/	fusion
iRBD vs sain	39%	39%	37%
MP vs sain	22%	22%	17%

TABLE 5.8 – EER issu de classification des hommes iRBD vs sains. à partir des enregistrements issus du microphone professionnel et comparaison avec la classification des hommes MP vs sain dans les mêmes conditions. Les tâches de lecture et répétition de phrases ont été testées à partir de GMM spécifiques, et la tâche /pataka/ à partir de GMM globaux. Pour la fusion le taux est obtenu en moyennant les deux scores de classification (lecture + répétition de phrases et /pataka/) pour chaque sujet.

Enfin nous avons effectué la même classification mais avec les enregistrements téléphoniques. Le nombre de participants différant légèrement nous avons choisi 30 iRBD et 30 sujets sains pour entrainer les GMM, et le reste (soit 7 iRBD et 6 sains) pour le test. Les résultats sont décrits dans le Tableau 5.9.

	repet	monologue	DDK
iRBD vs sain	40%	45 %	43 %
MP vs sain	36%	36 %	25 %

TABLE 5.9 – EER issus de classification des hommes iRBD vs sains, à partir des enregistrements téléphoniques, et comparaison avec la classification des hommes MP vs sain dans les mêmes conditions. Les tâches vocales répétition de phrase, monologue et DDK ont été testés par rapport à des GMM spécifiques. Toutes les sessions ont été utilisées pour l'entraînement et le test.

Tous nos participants iRBD ne sont pas au même stade d'évolution de la maladie vers un syndrome parkinsonien. Certains iRBD ont commencé à développer des symptômes moteurs, on les considère comme $moteur^+$ si leur score UPDRS III est supérieur ou égal à 14 (cf. partie 4). Parmi les iRBD hommes que nous avons classés précédemment, 18 sont $moteur^+$ et 23 $moteur^-$. Si on considère la classification de tous les iRBD hommes vs les sujets sains hommes, enregistrés avec le microphone professionnel, nous avions obtenu, en prenant le score de fusion, un EER de 37%. Si maintenant on cherche à classer seulement les iRBD $moteur^+$ par rapport aux sujets sains, nous obtenons un **EER de 28%.** A l'inverse si on cherche à classer les iRBD $moteur^-$ par rapport aux sujets sains, nous obtenons un EER de 41%.

Cette différence de performance semble montrer que le classifieur MFCC-GMM commence à détecter les iRBD quand ces derniers développent leurs premiers symptômes moteurs.

	iRBD $moteur^+$	iRBD $moteur^-$
fusion	28%	41%

TABLE 5.10 – EER issus de classification des hommes iRBD vs sains à partir des scores de fusion issus des enregistrements avec le microphone professionnel. Comparaison des performances des iRBD $moteur^+$ et $moteur^-$

5.3 Classification avec GMM-UBM

5.3.1 UBM à partir de nos données

Dans le but d'essayer d'améliorer les performances de classification, nous avons testé la technique de GMM-UBM présentée au paragraphe 3.3.1.3 consistant en une adaptation d'un modèle du monde avec nos données. Cette méthode est connue pour améliorer la reconnaissance du locuteur quand peu de données sont disponibles pour l'entraînement des modèles de locuteurs. D'après [Reynolds et al., 2000] l'amélioration des performances de classification étant due à l'augmentation des données servant à construire les GMM et au couplage induit entre les modèles. Afin d'abord de tester une amélioration potentielle due au couplage des modèles nous avons créé des modèles du monde à partir de nos sujets, en testant la méthode GMM-UBM sur la classification hommes MP vs hommes sains à partir des tâches lecture et répétitions de phrases enregistrées avec le microphone professionnel. Nous avons souhaité voir si cela améliorait notre baseline ayant un EER de 22%. Pour cela nous avons testé différentes façons de construire le modèle du monde et de l'adapter. Nous avons pris dans un premier temps un sous-groupe de nos sujets MP et sain pour entraîner l'UBM, que nous avons adapté à partir d'autres sujets MP et sains afin de construire un GMM MP et un GMM sain (Expérience 1). Dans un deuxième temps nous avons entrainé l'UBM à partir de MP, puis nous l'avons adapté à partir de sujets sains pour créer notre GMM sain (Expérience 2). Pour ces deux expériences, les modèles ont été testés sur des sujets tests différents de ceux ayant servi à l'UBM ou à l'adaptation. Dans un troisième temps nous avons pris tous les sujets pour entraîner l'UBM, puis nous l'avons adapté à partir de MP et de sains pour créer le GMM MP et le modèle sain, que nous avons testés sur les sujets restants (Expérience 3).

Voici le détail de la répartition des sujets pour ces trois expériences et l'EER du modèle agrégé obtenu. Nous avons considéré les tâches vocales de lecture et répétitions de phrase pour l'UBM, les adaptations et les tests. Nous avons utilisé respectivement 20, 20 et 50 gaussiennes pour créer les UBM.

Expérience 1 :
- UBM : 20 MP et 10 sains (45 min de données parole)
- GMM MP : adaptation MAP à partir de 30 MP (45 min)
- GMM sain : adaptation MAP à partir de 30 sains (45min)
- test : 24 MP et 8 sains (90s par test)
- EER= 22%

Expérience 2 :
- UBM : 36 MP (54 min)
- GMM MP : UBM
- GMM sain : adaptation MAP à partir de 36 sains (54min)
- test : 38 MP et 12 sains (90s par test)
- EER= 23%

Expérience 3 :
- UBM : 74 MP et 48 sains (3h)

- GMM MP : adaptation MAP à partir de 36 MP (54min)
- GMM sain : adaptation MAP à partir de 36 sains (54min)
- test : 38 MP et 12 sains (90s par test)
- EER= 22%

Nous constatons que le couplage des modèles induit par la méthode de GMM-UBM n'a pas amélioré les performances.

5.3.2 UBM à partir de données extérieures

Nous avons également testé l'impact d'une plus grande quantité de données en entrainant l'UBM sur des bases de données de conversations téléphoniques Switchboard collectées par le Linguistic Data Consortium (LDC) composées de 2594 sujets équilibrés en hommes et femmes, pour un total de 984 heures de données téléphoniques, échantillonnées en 8kHz [Graff et al., 1998, Graff et al., 1999, Graff et al., 2001, Graff et al., 2004]. Bien que la langue dans cette base soit l'anglais, elle devrait constituer un modèle du monde correct de par la ressemblance des phonèmes avec le français. L'utilisation d'un langage différent pour la construction des UBM a déjà été rencontrée lors d'évaluations de reconnaissance du locuteur SRE [Sadjadi et al., 2017].

Nous avons entrainé l'UBM à partir de ces données, échantillonnées en 8kHz, en prenant 2048 gaussiennes et une plage de [300-7000]Hz pour l'extraction des MFCC. Nous avons adapté cet UBM aux monologues issus de notre base téléphonique, ayant des caractéristiques (fréquences d'échantillonnage, plage de fréquence, microphones..) plus compatibles que les données enregistrées avec notre microphone professionnel.

Nous avons entrainé le modèle GMM MP en adaptant l'UBM à partir de 30 MP (soit 2h30 de données parole) et avons fait l'équivalent pour le GMM sain. Nous avons effectué les tests sur les sujets restants à savoir 33 MP et 6 sains (5min par test). Nous avons obtenu un EER de 36% (Expérience 4) soit pas d'amélioration par rapport à notre baseline sans UBM.

Nous avons également testé l'adaptation de l'UBM LDC à partir des tâches DDK (texte-dépendant), soit environ 2h30 de données pour l'adaptation du GMM MP et du GMM sain, toujours issues des données téléphoniques, et nous avons obtenu cette fois une dégradation de 3% (Expérience 5). Pour finir nous avons construit un UBM LDC à partir cette fois des sujets hommes uniquement, et nous l'avons adapté de la même manière que précédemment à nos sujets hommes MP et sains, sans obtenir d'amélioration des performances. La même chose a été faite pour les sujets femmes, à savoir l'UBM LDC à partir des sujets femmes et l'adaptation à partir 20 MP femmes et 20 sains femmes de notre base de données téléphonique, n'entrainant pas non plus d'amélioration des performances.

Expérience 4 :
- UBM : 2594 sujets (hommes et femmes) (984h) - tâche conversation
- GMM MP : adaptation MAP à partir de 30 MP hommes (2h30) - tâche monologue
- GMM sain : adaptation MAP à partir de 30 sains hommes (2h30) - tâche monologue
- test : 33 MP hommes et 6 sains hommes (5min par test) - tâche monologue
- EER= 36% à comparer à EER=36% sans UBM

Expérience 5 :
- UBM : 2594 sujets (hommes et femmes) (984h) - tâche conversation
- GMM MP : adaptation MAP à partir de 30 MP hommes (2h30) - tâche DDK
- GMM sain : adaptation MAP à partir de 30 sains hommes (2h30) - tâche DDK
- test : 33 MP hommes et 6 sains hommes (5min par test) - tâche DDK
- EER= 28% à comparer à EER=25% sans UBM

Nous pouvons donc conclure qu'entraîner un UBM à partir d'une grande base de données externe de type conversations téléphoniques puis l'adapter avec notre base de données téléphoniques n'améliore pas les performances de classification des tâches texte-indépendant et dégrade légèrement les performances des tâches texte-dépendant. Cette dégradation peut résulter du fait que les fichiers audio à partir desquels a été entrainé l'UBM sont des conversations donc par nature texte-indépendant et du coup moins compatibles avec une adaptation texte-dépendant. La non amélioration des performances issues de l'adaptation de l'UBM aux monologues de notre base de données par rapport à l'entraînement d'un GMM sans UBM semble indiquer que notre base de données est suffisamment grande pour la construction de GMM, ne nécessitant pas l'intervention d'une base extérieure pour ce type de classifieur.

5.4 Classification avec GMM sur les transitions "non voisé à voisé"

Récemment des études ont montré de bonnes performances de classification en s'intéressant aux transitions des sons voisés à non voisés et inversement [Orozco-Arroyave et al., 2015b, Vásquez-Correa et al., 2017a]. Les transitions "non voisé à voisé" correspondent à des attaques de sons et semblent être particulièrement discriminantes dans MP. Nous avons voulu tester si ces transitions pouvaient être utilisées seules pour construire un modèle GMM MP et un modèle sain à partir des MFCC, afin d'effectuer la classification. Pour cela nous avons utilisé une méthode d'autocorrélation à l'aide du logiciel Praat pour extraire la fréquence fondamentale, à raison d'environ une information sur la fréquence fondamentale toutes les 10ms. Cette information indique s'il existe une fréquence fondamentale pour la trame en question (donc si elle correspond à un son voisé) et si oui donne la valeur de cette fréquence. Ensuite nous avons extrait les temps de transitions séparant des séquences d'au moins 4 trames non voisées suivies d'au moins 4 trames voisées. A partir de ces temps de transitions nous avons pu isoler les morceaux des monologues correspondant à ces transitions. Nous avons choisi d'utiliser des morceaux d'une durée de 60ms, centrés autour des temps de transitions (soit 30ms de sons non voisés suivis de 30ms de sons voisés, faisant un total de 6 vecteurs MFCC). Nous avons utilisé la même méthode que partie 5.1 pour extraire les MFCC et effectuer la classification. Nous avons entrainé les GMM à partir des transitions extraites des monologues de 36 MP hommes et 36 sains hommes, et testé les sujets restants. Ceci représentait environ 3min de données paroles par GMM et 6s par sujet test. Nous avons utilisé 20 gaussiennes pour les GMM et à la différence de partie 5.1 nous avons choisi de calculer les deltas des MFCC sur 2 trames consécutives au lieu de 3, vu la courte durée des extraits audio correspondant aux transitions. Nous avons obtenu un EER de 40%, ce qui signifie que la méthode MFCC-GMM n'est pas la meilleure méthode pour détecter MP à partir des transitions non voisé à voisé issues du monologue.

Dans le but d'améliorer la classification à partir des transitions non voisé à voisé, nous avons choisi de garder seulement les transitions correspondant au phonème /p/. En effet comme nous l'avions expliqué section 2.3.2, l'articulation des consonnes occlusives est un des éléments les plus touchés dans la dysarthrie parkinsonienne. Donc on peut supposer que les transitions non voisé à voisé correspondant à des attaques de consonnes occlusives contiennent plus d'informations discriminantes que les autres. Nous avons alors choisi d'extraire les transitions non voisé à voisé à partir de la tâche de répétitions lentes de syllabes /pa/. Nous avons utilisé la même méthode que précédemment pour isoler les transitions, extraire les MFCC et effectuer la classification. Ce qui représentait en moyenne 1min de données paroles pour chacun des GMM MP et sain, et 2s par sujet test. Nous avons obtenu un EER de $27\% \pm 6\%$, cf. Tableau 5.11, ce qui est un bon résultat vu la faible quantité de données utilisées pour l'entraînement et le test.

Pour la comparaison nous avons également entrainé les GMM et effectué une classification

à partir de la totalité de la tâche de répétition lente /pa/, et plus seulement à partir des transitions non voisé à voisé, ce qui a conduit à un EER de 29%. Donc le fait d'isoler les attaques des phonèmes /p/, améliore de 2% les performances par rapport aux syllabes /pa/ entières.

Nous avons aussi effectué une classification à partir des transitions non voisé à voisé extraites de la tâche de répétitions rapides de syllabe /pa/ et obtenu cette fois un EER de 41%. Cette détérioration importante des performances peut s'expliquer par le fait que la détection des transitions non voisé à voisé est plus difficile lors des répétitions rapides : on constate que pour certains sujets presque toutes les trames sont considérées comme voisées. Pour faciliter la détection des phonèmes non voisés on a essayé par la suite d'abaisser le seuil de voisement, mais cela a entrainé la perte des vrais sons voisés chez d'autres sujets.

tâches	durée GMM	durée test	EER
monologue	3min	6s	40%
/pa/ lent	1min	2s	27%
/pa/ rapide	30s	1s	41%

TABLE 5.11 – Résultats de la classification MP hommes vs sains hommes à partir des transitions non voisé à voisé issues des enregistrements faits avec le microphone professionnel. Pour le monologue les transitions correspondent à l'attaque des sons voisés et pour les tâches de répétition des syllabes /pa/ rapides (DDK) et lentes, les transitions correspondent à la prononciation du phonème /p/. Les durées GMM correspondent à la durée des données voix utilisées pour l'entraînement du GMM MP et du GMM sain, composé de 36 sujets chacun. Les durées test correspondent à la durée des données voix utilisées pour le test de chaque sujet.

La classification des femmes par rapport au phonème /p/ a quant à elle conduit à un EER de 49% soit aucune différence avec le hasard. Concernant la classification des iRBD par rapport aux sujets sains, nous avons obtenu un EER de 43%. Ces diminutions de performance concernant les femmes et les iRBD sont du même ordre que partie 5.1 en prenant les tâches entières.

Pour finir, afin d'évaluer l'impact des conditions d'enregistrement sur la classification MFCC-GMM à partir des phonèmes /p/ issus de tâches de répétions lentes de la syllabe /pa/, nous avons effectué également une classification MP hommes vs sains hommes à partir des enregistrements du microphone interne de l'ordinateur (avec l'option de débruitage actif) et des enregistrements téléphoniques (toutes sessions confondues). Nous avons obtenu un EER de 42% pour ces deux bases de données. Ces détériorations de performances sont cohérentes avec les résultats précédents. Les enregistrements issus du microphone interne de l'ordinateur conduisent à de moins bonnes performances pour les tâches DDK et les phonèmes /p/, semblant ainsi masquer une partie des différences MP vs sains que l'on peut observer dans la prononciation des consonnes occlusives. Ceci pouvant être dû à la distance accrue entre le sujet et le microphone, à la fonction de réduction de bruit active ou à la moins bonne qualité du microphone. Quant aux enregistrements téléphoniques, le fait de cumuler toutes les sessions, procurant 6s de données parole par sujet en moyenne, n'entraine pas une augmentation suffisamment importante de la taille de la base de données pour pallier la qualité réduite des enregistrements.

Pour conclure, en extrayant seulement les phonèmes /p/ de la tâche de répétition lente de syllabes /pa/ enregistrée avec le microphone professionnel, soit environ 2s de données parole par sujet, nous avons pu classer les MP hommes par rapport aux sains hommes avec un EER de 27%. Même si ce résultat n'est atteint que pour les hommes et avec le microphone de bonne qualité, il va dans le sens d'un pouvoir discriminant important du phonème /p/ dans la détection de MP.

5.5 Conclusion sur les analyses MFCC-GMM

En résumé, cette analyse a consisté à adapter une méthode utilisée en reconnaissance du locuteur pour la détection de MP débutant, à partir de différents types de tâches vocales, et différents types d'enregistrements (avec un microphone professionnel, un microphone interne d'ordinateur et l'usage de téléphones). Nous avons fait les analyses séparément pour les hommes et pour les femmes, afin de ne pas rajouter la variabilité due au genre.

La première étape a consisté en divers prétraitements (comme la soustraction spectrale), afin entre autres de supprimer l'effet du non appariement complet de l'environnement acoustique entre les groupes, dû aux différents lieux d'enregistrement.

Nous avons ensuite extrait 19 MFCC, la log énergie, les Deltas et les Deltas-deltas, toutes les 10ms. Afin de ne garder que les trames sonores, nous avons effectué une VAD, et pour limiter l'effet du canal, nous avons fait une CMS.

Nous avons séparé les sujets en groupes d'entraînements (un groupe MP et un groupe sain) et groupe de test. A partir des groupes d'entraînement, nous avons construit un modèle GMM MP et un modèle GMM sain, décrivant la distribution des MFCC des MP et des sains. Nous avons ensuite testé la vraissemblance (LLH) des vecteurs MFCC des sujets tests par rapport aux deux modèles GMM et calculé un score compris entre 0 et 1 à partir des LLH ratios.

Pour la construction du modèle final nous avons utilisé une méthode ensembliste de type *repeated random subsampling aggregation*, dont la performance a été évaluée en moyennant les scores des sujets correspondants à chaque run où ils ne faisaient pas partie du groupe d'entraînement. L'EER et les courbes DET ont été ensuite calculés pour comparer les performances liées aux différentes tâches et aux différents types de microphones utilisés.

Les meilleures performances ont été obtenues à partir des tâches de lecture et répétitions de phrases ainsi qu'à partir des tâches DDK (la tâche /pataka/ étant la plus performante). La tâche de type monologue s'est révélée un peu moins efficace. Ceci peut s'expliquer par les différences de contenu phonétique d'un sujet à l'autre, apportées par cette tâche texte-indépendant, qui créent une variabilité pouvant diminuer le pouvoir discriminant des modèles. Enfin les répétitions lentes et surtout les voyelles soutenues se sont révélées peu appropriées pour ce type d'analyse.

Nous avons également évalué l'influence du contenu des données utilisées pour l'entraînement au regard des tâches utilisées pour le test. Nous sommes arrivés à la conclusion que pour le test des tâches texte-dépendant comme la lecture et les répétitions de phrases, il était préférable d'avoir entrainé les GMM de façon spécifique (c'est à dire avec ces mêmes tâches) plutôt qu'avec toutes les tâches. Ceci étant valable à condition d'avoir assez de quantité de paroles pour entraîner correctement les GMM à partir des tâches en question. Dans le cas contraire, rencontré par exemple si on considère seulement les tâches /pataka/, il est préférable d'augmenter la quantité de données d'entraînement, quitte à rendre les GMM moins spécifiques. Ainsi le choix des données utilisées pour l'entraînement des GMM résulte d'un compromis entre quantité et spécificité.

La fusion des deux meilleures tâches, à savoir la lecture + répétition de phrase testées par rapport à des GMM spécifiques, et la tâche /pataka/ testée par rapport à des GMM globaux, conduit à un EER de 17% chez les hommes, pour les enregistrements effectués avec le microphone professionnel.

La classification MP vs sain pour les femmes s'est révélée beaucoup moins performante que pour les hommes avec ce type d'analyse (EER de 42%). Ceci peut être la conséquence de la plus grande variabilité des MFCC chez les femmes rendant les classifications par les méthodes MFCC-GMM plus compliquées [Fraile et al., 2009b]. Ceci pourrait également confirmer l'hypothèse d'une atteinte neuronale moindre, ou au moins différente, au début de la maladie de Parkinson chez les femmes [Haaxma et al., 2007].

Concernant les iRBD, pouvant être considérés comme au stade prodromique de la MP, nous avons pu les classer par rapport aux sujets sains avec un EER de 37%. Si on considère seulement les iRBD $moteur^+$, cette EER descend à 28%, ce qui semble indiquer que cette méthode d'analyse décèle les pré-parkinsoniens à partir du moment où ils commencent à avoir quelques perturbations motrices (mais pas suffisantes pour poser le diagnostic de MP).

Concernant l'influence du type de microphone utilisé, nous avons pu comparer les résultats du microphone professionnel avec ceux obtenus à partir des enregistrements simultanés du microphone interne de l'ordinateur. Nous avons observé une dégradation moyenne de 8% avec le microphone de l'ordinateur, lors de la classification des hommes MP vs sain et RBD vs sain. Cette dégradation étant légèrement moins importante pour la lecture mais plus importante pour les tâches DDK. Les causes de cette dégradation sont sans doute liées à la distance accrue entre la bouche et le microphone (15cm au lieu de 5cm) et à la fonction de débruitage actif de l'ordinateur. La qualité réduite du microphone de l'ordinateur pouvant également contribuer légèrement à cette dégradation.

Nous avons également testé l'effet "cross-micro" en utilisant nos données issues du microphone professionnel pour l'entraînement des modèles et en testant les sujets enregistrés avec le microphone de l'ordinateur. Nous avons obtenu un EER de 25% pour la lecture et la répétition de phrases. Ce résultat donne l'ordre de grandeur des performances que l'on pourrait avoir en testant des enregistrements effectués dans d'autres conditions (avec d'autres microphones) à partir de nos modèles.

Concernant les enregistrements téléphoniques, nous avons constaté une dégradation des performances de 18%, pour la classification des hommes MP vs sain, en utilisant une session téléphonique par sujet test, l'entraînement des GMM étant fait à partir de toutes les sessions des sujets d'entraînement. Afin de mieux comprendre les causes de cette dégradation nous avons réalisé une simulation grossière du téléphone, à partir de nos enregistrements issus du microphone professionnel, en sous échantillonnant à 8kHz et en considérant la bande de fréquence étroite liée aux transmissions téléphoniques. Nous avons alors obtenu une dégradation des performances de 9% par rapport au microphone professionnel. Ce résultat indique que notre détérioration de 18% résulte pour moitié de l'échantillonnage plus faible et de la bande de fréquence étroite. L'autre moitié serait la conséquences des autres caractéristiques du téléphone, comme le bruit, la distorsion due aux codecs.. ainsi que la qualité réduite d'exécution des tâches qui sont réalisées en autonomie.

Quand cette fois toutes les sessions téléphoniques sont utilisées pour le test de chaque sujet, nous constatons une amélioration de 10%, en comparaison aux performances obtenues en utilisant une seule session téléphonique par sujet test, soit un EER de 25% pour la détection des hommes MP vs sain. Cette amélioration est obtenue dès 5 sessions téléphoniques, soit à partir de 5min de données DDK par sujet.

Pour vérifier la pertinence du modèle agrégé par rapport au modèle simple, nous avons également calculé l'EER du modèle simple (par la méthode standard de validation croisée random subsampling). Nous avons obtenu une diminution des performances de 2 à 3% pour le modèle simple, que ce soit avec les données du microphone professionnel ou avec les données téléphoniques, confirmant l'intérêt de la méthode ensembliste.

Dans le but de tester l'effet de l'apport d'une plus grande quantité de données pour la construction des GMM, nous avons testé la méthode GMM-UBM en entrainant nos GMM à

partir d'un modèle UBM, entrainé avec une grande base de données voix publique, et que nous avons adapté (adaptation MAP) avec nos données d'entraînement. Afin d'évaluer le simple effet de couplage induit par cette méthode nous avons également construit des UBM à partir de nos données. Dans les deux cas, nous n'avons pas observé d'amélioration de performances.

Vu les travaux récents [Orozco-Arroyave et al., 2015b] montrant l'intérêt des analyses sur les transitions des sons non voisés à voisés, nous avons également effectué notre classification MFCC-GMM à partir des attaques des sons voisés du monologue et des tâches de répétition rapides et lentes des syllabes /pa/. Nous avons constaté que les attaques des occlusives /p/ étaient spécialement discriminantes dans la maladie de Parkinson, car une classification seulement à partir de ces sons a conduit à un EER de 27% (avec seulement l'équivalent de 2s de données par sujet testé).

Pour conclure, la méthode d'analyse MFCC-GMM s'est avérée pertinente pour la classification des hommes MP vs sain, avec un EER de 17% avec les enregistrements du microphone professionnel et 25% avec les enregistrements téléphoniques. Son intérêt est d'autant plus grand que cette méthode nécessite peu de données et que son coût computationnel est faible. Des améliorations de performances sont possibles, surtout pour les femmes (pour lesquelles la variabilité des MFCC semble rendre cette méthode inefficace), avec des méthodes de classifications plus modernes et plus coûteuses computationnellement. Il pourrait être intéressant de tester des méthodes utilisant les supervecteurs, les i-vecteurs, ou des méthodes spécialement spécifiques aux tâches texte-dépendant comme les Hidden Markov Models (HMM) et les réseaux de neurones de types Recurrent Neural Network (RNN).

Dans l'analyse suivante, nous avons adapté la dernière méthode en date utilisée en reconnaissance du locuteur, dont les performances dépassent celles des GMM dans ce domaine, mais nécessitant beaucoup de données et étant plus coûteuse computationnellement. C'est la première fois que cette méthode est utilisée dans le cadre de la détection de MP.

Chapitre 6

Classification MP vs sain à partir des x-vecteurs

Nous avons souhaité comparer les résultats obtenus avec la méthode MFCC-GMM avec la méthode la plus récente utilisée en reconnaissance du locuteur, décrite partie 3.3.1.3, qui est celle des x-vecteurs. Pour cela nous avons utilisé des DNN pré-entrainés dans le cadre d'une reconnaissance du locuteur, nous avons extrait les x-vecteurs de chaque sujet et avons effectué la classification MP vs sain en comparant le x-vecteur d'un sujet test à la moyenne des x-vecteurs MP et à la moyenne des x-vecteurs sains, calculés sur des groupes d'entraînement dans lesquels ne faisait pas partie le sujet test. Pour la comparaison des x-vecteurs et donc l'étape de classification, nous avons comparé différentes techniques : distance cosinus sans et après LDA, et PLDA. Nous avons également étudié l'impact de la longueur des segments tests, ainsi que l'influence de l'augmentation de données sur la LDA et la PLDA. Pour finir nous avons effectué une dernière comparaison avec un DNN entrainé avec nos propres données. Comme pour les GMM, nous avons considéré les hommes et les femmes séparément et avons utilisé une méthode ensembliste pour le résultat de classification final. Comme les classifications par x-vecteur + LDA ou PLDA ont besoin de beaucoup de données pour l'entraînement, nous avons commencé par analyser ces méthodes sur nos données téléphoniques (plus nombreuses), avec le DNN pré-entrainé SRE16 (8kHz). Les données issues du microphone professionnel ont été analysées dans un deuxième temps, l'extraction des x-vecteurs ayant été faite à partir du DNN pré-entrainé voxceleb (16kHz).

6.1 Méthode

6.1.1 Extraction des MFCC

Pour les caractéristiques concernant la fréquence d'échantillonnage, les MFCC et la VAD, nous avons gardé les mêmes que celles utilisées pour l'entraînement des DNN.

6.1.1.1 Analyse téléphone

Pour l'analyse des enregistrements téléphoniques, échantillonnés en 8kHz, nous avons extrait 23 MFCC et la log énergie toutes les 10ms sur des fenêtres de 25ms. La bande de fréquence utilisée pour le calcul des MFCC est de [20-3700Hz]. Tout comme pour les GMM, afin de supprimer l'effet convolutif du canal, une soustraction de cepstre moyen est calculée sur des fenêtres glissantes de 3s. Une VAD identique à celle décrite partie 5.1.2 est ensuite effectuée. Contrairement à notre analyse GMM, nous n'avons pas calculé les delta ni delta delta MFCC car la composante temporelle est déjà prise en compte dans les premières couches du DNN.

6.1.1.2 Analyse microphone professionnel

Pour l'analyse de nos enregistrements issus du microphone de haute qualité, nous avons sous échantillonné les enregistrements en 16kHz, et extrait 30 MFCC plus la log énergie toutes les 10ms sur des fenêtres de 25ms. La bande de fréquence utilisée pour le calcul des MFCC est de [20-7600Hz]. De la même manière que pour les analyses téléphoniques, une CMS et une VAD avec les mêmes paramètres ont été effectuées.

Tout comme pour les analyses MFCC-GMM, l'étape d'extraction des MFCC, pour les enregistrements du microphone professionnel, a été précédée d'une étape de prétraitement par soustraction spectrale, afin de supprimer le biais dû au non appariement de l'environnement sonore.

6.1.2 Entraînement DNN

Nous avons utilisé des DNN préentrainés (modèle SRE16 et modèle voxceleb) pour la reconnaissance du locuteur et disponibles en ligne ($http://kaldi-asr.org/models.html$).

Le modèle SRE16, décrit dans [Snyder et al., 2018b], que nous avons utilisé pour l'extraction des x-vecteurs de données téléphoniques, a été entrainé sur 5139 sujets. Les bases de données utilisées appartiennent au catalogue LDC, et comprennent les corpus Swichboard (Phase1,2,3 et Cellular 1,2), Mixer 6 et des évaluations NIST SREs. Ces bases de données regroupent un mélange de conversations téléphoniques et de données enregistrées avec microphone, avec comme langue dominante l'anglais. Certaines données sont directement échantillonnées en 8kHz, et les enregistrements échantillonnés en 16kHz sont alors sous échantillonnés en 8kHz. Enfin ces données ont été augmentées avec les bases MUSAN ($http://www.openslr.org/17$) et RIR NOISES ($http://www.openslr.org/28$) sous-échantillonnées à 8kHz (cf. partie 6.1.6).

Le modèle voxceleb, que nous avons utilisé pour l'extraction des x-vecteurs de données téléphoniques, a été entrainé sur la base voxeleb [Nagrani et al., 2017]. Les données proviennent d'interview video de 7330 célébrités postées sur Youtube. La voix est échantillonnée à 16kHz. Les données ont été augmentées avec les bases MUSAN et RIR NOISES.

Ces deux modèles sont des réseaux de neurones profonds (DNN) présentant l'architecture introduite partie 3.3.1.3 et détaillée Tableau 6.1. La partie TDNN est composée de 5 couches *frame-level* prenant comme entrée les vecteurs MFCC avec un contexte temporel venant des vecteurs MFCC des trames voisines. L'étape de pooling rassemble les outputs du TDNN pour toutes les trames du segment input en calculant les moyennes et écarts types. Enfin la dernière partie du réseau est constituée de deux couches *segment-level* prenant comme entrée les statistiques du segment issues de l'étape de pooling, et une couche finale softmax. Les x-vecteurs sont extraits après la première couche *segment-level* juste avant la fonction d'activation non linéaire ReLu.

Pour l'expérience dont les résultats sont présentés partie 6.2.1.5, nous avons entrainé un DNN avec la même architecture mais avec nos données. La seule différence d'architecture est la dimension de l'output sortant de la couche softmax qui est alors de deux. Les sujets utilisés pour l'entraînement du DNN sont alors les mêmes que ceux utilisés pour la constitution des x-vecteur moyens sain et MP et pour l'entraînement de la LDA et PLDA. La classification étant effectuée sur des sujets différents (appartenant au groupe test).

6.1.3 Extraction des x-vecteurs

Les x-vecteurs sont extraits pour chaque segment à partir des DNN pré-entrainés. Ce sont des vecteurs de dimension 512 et sont une représentation du segment voix pris comme input. Les segments utilisés pour l'entraînement des DNN ont une durée de [2-4s] (une fois les silences enlevés). Pour l'extraction d'x-vecteurs à partir de nouvelles données, les durées compatibles avec ces DNN pré-entrainés vont de 25ms à 100s. C'est à dire que les segments audio d'une durée

couche	nombre trames	dimension input	dimension output
frame-level 1	5	5*K	512
frame-level 2	9	1536	512
frame-level 3	15	1536	512
frame-level 4	15	512	512
frame-level 5	15	512	1500
pooling	T	1500*T	3000
segment-level 6	T	3000	512
segment-level 7	T	512	512
softmax	T	512	N

TABLE 6.1 – Architecture des DNN ayant servi à l'extraction des x-vecteurs. Ces derniers sont extraits au niveau de la couche 6 avant l'application de la fonction d'activation ReLu. T est le nombre de trames contenues dans le segment. K est le nombre de paramètres pour une trame, soit 24 pour les enregistrements téléphoniques (23MFCC + log énergie) et 31 pour les enregistrements professionnels (30MFCC + log énergie). N est le nombre de locuteurs ayant servi pour l'entraînement, soit 5139 pour le modèle SRE16 et 7330 pour le modèle voxceleb.

inférieure à 25ms ne sont pas pris en compte. Concernant les segments d'une durée supérieure à 100s, ils sont découpés en morceaux inférieurs à 100s. Les x-vecteurs sont extraits pour chaque morceau puis moyennés.

Nous avons dans un premier temps segmenté les fichiers audio de notre base téléphonique d'entraînement en segments de [1-5s], afin d'avoir suffisamment de x-vecteurs pour entraîner la LDA et la PLDA. Pour les données de test nous avons utilisé un fichier audio par personne et par type de tâche. Les fichiers supérieurs à 100s étant alors redécoupés pour l'extraction des x-vecteurs comme expliqués ci-dessus. De fait les segments testés allaient de 15s à 100s, donc d'une durée supérieure à ceux utilisés pour l'entraînement. Nous avons comparé les résultats obtenus avec ceux obtenus en segmentant des fichiers tests de la même manière que pour l'entraînement, donc avec des durées comparables de [1-5s] Les différents résultats des performances de classifications obtenues sont présentés partie 6.2.1.1. Par la suite nous avons gardé la deuxième méthode avec les segments de tests et d'entraînement de même durée.

6.1.4 Comparaison des x-vecteurs

Une fois les x-vecteurs extraits pour chaque sujet, les x-vecteurs des sujets appartenant au groupes d'entraînement MP et sain sont moyennés de manière à avoir un x-vecteur MP et un x-vecteur sain de référence. La classification des sujets test s'opère en comparant leur x-vecteur au x-vecteurs MP et sain, la différence entre ces deux "distances" est ensuite calculée puis normalisée avec une fonction sigmoïde. Dans le cas où les fichiers audio des sujets tests sont segmentés en plusieurs parties, un x-vecteur est extrait par segment et la comparaison a alors lieu pour chaque segment. La moyenne des scores de classification sur tous les segments donne alors la classe du sujet. Plusieurs méthodes existent pour mesurer la distance entre des vecteurs. Nous avons comparé 3 méthodes souvent utilisées avec les i-vecteurs ou les x-vecteurs : la distance cosinus, la distance cosinus précédée d'une LDA et la PLDA. Ces trois méthodes sont présentées partie 3.3.1.3. Nous avons entrainé une LDA à 2 dimensions (car nous avons 2 classes) à partir des x-vecteurs des sujets d'entraînement labellisés. Pour la PLDA nous l'avons précédée d'une LDA, et avons utilisé les mêmes sujets d'entraînement.

6.1.5 Classification finale et validation

Pour la classification finale et la validation nous avons gardé la méthode ensembliste présentée précédemment partie 5.1.3.3. Nous avons effectué 40 runs de classification, pour lesquels les participants étaient répartis de manière aléatoire entre les groupes d'entraînement et de test (repeated random subsampling). Puis nous avons effectué une agrégation des scores pour chaque sujet, ne moyennant les scores de classification des runs où ils ont été testés. Concernant la répartition des sujets entre les groupes d'entraînement et de test, nous avons gardé les mêmes proportions que celles utilisées précédemment avec l'analyse GMM.

6.1.6 Augmentation de données

Récemment des améliorations dans la reconnaissance du locuteur avec les i-vecteurs et les x-vecteurs ont été trouvées en augmentant les données [Snyder et al., 2018b]. L'augmentation consiste à dupliquer les données en y rajoutant des bruits additifs et de l'écho, augmentant ainsi la quantité et la diversité des échantillons utilisés pour l'entraînement du DNN et de la PLDA. A la manière de [Snyder et al., 2018b], nous avons utilisé 4 types différents d'augmentation des données :

- Echo : Une simulation d'écho est effectuée sur les données voix à partir de la convolution des données avec une réponse impulsionnelle de salles (RIR pour *Room Impulse Responses*) de différentes formes et tailles disponibles en ligne (*http* : *//www.openslr.org/*28).

- Bruit additif : différents types de bruits (extraits à partir de la base de données MUSAN, *http* : *//www.openslr.org/*17), sont ajoutés de manière additive, toutes les secondes.

- Musique additive : des extraits musicaux (issus de la base MUSAN) sont ajoutés en bruit de fond aux données.

- Brouhaha additif : des données voix de 3 à 7 locuteurs (issus de la base MUSAN) sont sélectionnées au hasard, et additionnées entre elles, créant un brouhaha ajouté à nos données voix en bruit de fond.

La moitié des 4 copies augmentées est alors choisie au hasard et ajoutée à notre base de données pour l'entraînement de la LDA et la PLDA, multipliant ainsi par trois la taille de cette dernière.

6.2 Résultats

6.2.1 Classification des hommes avec le téléphone

Nous avons commencé par classer les hommes MP par rapport aux hommes sains avec les données téléphoniques. De la même manière qu'avec les GMM, 30 hommes MP et 30 hommes sains ont été utilisés à chaque run pour l'entraînement (et la constitution des x-vecteurs de référence MP et sain) et 33 hommes MP et 6 hommes sains pour le test. Toutes les sessions analysables de ces sujets ont été utilisées. Pour le monologue cela fait un total moyen de 5h de données voix pour le groupe d'entraînement à chaque run, et un total de 5min par sujet test. Les mêmes quantités de voix sont utilisées pour les tâches DDK. Pour les répétitions de phrases, cela revient à environ 1h30 pour le groupe d'entraînement et 1min30 par sujet test. Le DNN pré-entrainé utilisé est le modèle SRE16 décrit dans la partie méthode précédente.

6.2.1.1 Influence de la durée des segments tests

Nous avons dans un premier temps classé nos sujets en utilisant un fichier audio par sujet pour le test. Ce qui signifiait, comme expliqué dans la méthode, des x-vecteurs extraits sur des segments de 15 à 100s pour le test. Sachant que la longueur des segments utilisés pour l'entraînement de la LDA et PLDA et la constitution des x-vecteurs de références MP et sain était de 1 à 5s, et celle pour l'entraînement du DNN de 2 à 4s. Les résultats obtenus avec la classification par distance cosinus, sans et avec LDA, et par PLDA sont présentés Tableau 6.2. Afin d'évaluer l'effet du non appariement de la longueur des segments entre entraînement et test, nous avons refait l'expérience en segmentant les fichiers tests de la même manière que pour l'entraînement (à savoir en segments de 1 à 5s). Les résultats obtenus sont détaillés Tableau 6.3. Pour ces deux premières expériences, nous n'avons pas effectué d'augmentation de données pour l'entraînement de la LDA et PLDA. Les données ayant servi à l'entraînement du DNN SRE16 avait par contre elles été augmentées.

Tâche	dist cos	LDA + dist cos	PLDA
repet	41 %	36 %	36 %
monologue	36 %	37 %	35 %

TABLE 6.2 – EER issus de la classification des hommes MP vs sain, à partir des tâches de répétitions de phrases et du monologue des enregistrements téléphoniques. Les x-vecteurs ont été extraits à partir de segments d'une durée de [1-5s] pour l'entraînement et [15-100s] pour le test.

Tâche	dist cos	LDA + dist cos	PLDA
repet	39 %	32 %	33 %
monologue	33 %	35 %	36 %

TABLE 6.3 – EER issus de la classification MP hommes vs sains hommes, à partir des tâches de répétitions de phrases et du monologue des enregistrements téléphoniques. Les x-vecteurs ont été extraits à partir de segments d'une durée de [1-5s] pour l'entraînement et le test.

Nous pouvons constater une amélioration globale de l'ordre de 2 à 3% lorsque les segments de test font la même longueur que les segments d'entraînement.

6.2.1.2 Influence de l'augmentation de données

Nous avons ensuite voulu tester l'effet de l'augmentation de données pour l'entraînement de la LDA et PLDA. Les résultats sont présentés dans le Tableau 6.4.

Tâche	LDA + dist cos	PLDA
repet	33 %	31 %
monologue	33 %	33 %

TABLE 6.4 – EER issus de la classification MP hommes vs sains hommes, à partir des tâches de répétitions de phrases et du monologue des enregistrements téléphoniques. Une augmentation de données a eu lieu pour l'entraînement de la LDA et la PLDA.

Nous pouvons constater une amélioration des performances pour la tâche de monologue mais pas pour la tâche de répétition. Ceci pouvant s'expliquer par le fait que l'augmentation de données introduit de la variabilité phonétique pouvant ainsi nuire aux tâches dépendantes du texte qui ont l'avantage de présenter un contenu phonétique similaire d'un sujet à l'autre.

6.2.1.3 Comparaison du modèle agrégé vs modèle simple

De la même manière que pour l'analyse GMM, nous avons souhaité comparer les performances de notre modèle de classification agrégé avec celles qu'on aurait avec le modèle simple. Pour estimer les performances du modèle simple nous avons moyenné les courbes DET issues de chaque run, et calculé l'EER correspondant à la courbe DET moyenne. Les performances obtenues sont détaillées Tableau 6.5. Nous pouvons constater une amélioration d'environ 2% pour le modèle agrégé, comparé au modèle simple. Cette amélioration due à la méthode ensembliste est du même ordre que celle constatée avec les analyses GMM (cf. partie 5.2.1.6).

	Tâche	LDA + dist cos	PLDA
modèle simple	repet	35 %	35 %
	monologue	35 %	35 %
modèle agrégé	repet	32 %	33 %
	monologue	33 %	33 %

TABLE 6.5 – Comparaison des EER du modèle simple et du modèle agrégé. Classification MP hommes vs sains hommes à partir des tâches de répétitions de phrases et du monologue des enregistrements téléphoniques. Une augmentation de données a eu lieu pour l'entraînement de la LDA et la PLDA du monologue.

6.2.1.4 Cas de la tâche DDK

Après avoir analysé les tâches de monologue et répétition de phrases avec les x-vecteurs, nous avons réalisé la classification à partir des tâches DDK, sans puis avec augmentation de données pour la LDA et PLDA. Comme pour la tâche de répétition de phrases, nous constatons que l'augmentation de données n'améliore pas les performances, ce qui est cohérent avec le fait que les tâches DDK sont également texte-dépendant.

Nous observons une dégradation des performances par rapport aux résultats obtenus avec l'analyse MFCC-GMM (EER=25%). Cette détérioration peut être expliquée par le fait que le DNN a été entrainé avec de la parole issue principalement de conversations, incluant une large variété de phonèmes. Les tâches DDK ne font intervenir qu'un nombre restreint de phonèmes et articulés dans un certain ordre. La spécificité de ces tâches n'est pas exploitée par la calibration du DNN, résultant en une perte du pouvoir discriminant par rapport aux GMM.

Tâche	distance cos	LDA + dist cos	PLDA
DDK non augmenté	35 %	29 %	30 %
DDK augmenté	-	30 %	30 %

TABLE 6.6 – EER issus de la classification MP hommes vs sains hommes, à partir des tâches de diadococinésie des enregistrements téléphoniques. Comparaison avec et sans augmentation de données pour LDA et PLDA.

6.2.1.5 Entraînement DNN avec notre base de données

Afin de rendre le DNN plus adapté pour le type particulier des tâches DDK, nous avons fait un essai en entrainant le DNN nous-mêmes avec nos données. Nous avons utilisé pour chaque run les données du groupe d'entraînement pour entraîner le DNN, avec une augmentation de données. Les résultats obtenus sont présentés Tableau 6.7. Nous constatons une dégradation des performances quand on effectue l'augmentation de données pour l'entraînement de la LDA et

PLDA. Les résultats issus de la classification par distance cosinus avec LDA et PLDA sans augmentation de données, sont similaires à ceux obtenus avec le DNN pré-entrainé. L'entraînement de notre DNN est certes plus spécifique mais souffre peut-être du manque de données, ce qui pourrait expliquer pourquoi il ne conduit pas à de meilleures performances.

Tâche	distance cos	LDA + dist cos	PLDA
DDK non augmenté	47 %	29 %	30 %
DDK augmenté	-	39 %	38 %

TABLE 6.7 – EER issus de la classification MP hommes vs sains hommes, à partir des tâches de diadococinésie des enregistrements téléphoniques. X-vecteurs extraits à partir d'un DNN entrainé avec notre base de données. Comparaison avec et sans augmentation de données pour LDA et PLDA.

6.2.1.6 Synthèse

Nous avons synthétisé les résultats obtenus lors de classification homme MP vs homme sain dans le Tableau 6.8 en comparant les performances des trois méthodes que nous avons utilisées pour la classification à partir des x-vecteurs (distance cosinus sans et avec LDA, et PLDA) avec les résultats obtenus lors de la classification MFCC-GMM pour les tâches monologue, répétition de phrase et DDK.

Nous constatons globalement que l'ajout de la LDA avant de calculer la distance cosinus améliore significativement les résultats, et sont au même niveau que la PLDA. Nous observons que l'augmentation de données améliore bien les résultats de la tâche texte-indépendant mais pas des tâches texte-dépendant. Ainsi le classifieur le plus approprié pour la détection de MP à partir des x-vecteurs semble être la distance cosinus précédé de la LDA (il n'y a pas d'utilité de garder la PLDA, plus complexe, si elle n'améliore pas la résultats), entrainée avec augmentation de données seulement pour le monologue. Le Tableau résumé 6.9 permet de comparer aisément les résultats issus de l'analyse MFCC-GMM avec la classification LDA + distance cosinus à partir des x-vecteurs. Nous observons que cette dernière est plus performante de 3% que la première pour les tâches de paroles "classiques" que sont le monologue et la répétition de phrases, mais moins performante de 5% pour les tâches DDK plus spécifiques.

Tâche	MFCC-GMM	dist cos	LDA + dist cos	PLDA	LDA aug + dist cos	PLDA aug
repet	35 %	39 %	**32** %	33 %	33 %	31 %
monol	36 %	33 %	35 %	36 %	**33** %	33 %
DDK	25 %	35 %	**29** %	30 %	30 %	30 %

TABLE 6.8 – EER issus de la classification MP hommes vs sains hommes, à partir des tâches de répétitions de phrase, monologue et diadococinésie des enregistrements téléphoniques. Comparaison types de classifications et effet augmentation de données.

6.2.2 Classification des femmes avec le téléphone

Nous avons ensuite effectué les mêmes types de classification mais cette fois pour les femmes, toujours avec les données téléphoniques. Pour chaque run, 20 femmes MP et 20 femmes saines ont été utilisées pour l'entraînement, et la constitution des x-vecteurs MP et sain de références. Le reste des sujets, à savoir 18 femmes MP et 5 femmes saines, figuraient dans le groupe test. Pour le monologue cela fait un total moyen de 3h20 de données voix pour le groupe d'entraînement à chaque run, et un total de 5min par sujet test. Pour les répétitions de phrases, cela revient à environ 1h pour le groupe d'entraînement et 1min30 par sujet test.

Tâche	MFCC-GMM	LDA + dist cos
repet	35 ± 8%	**32** ± 8%
monologue	36 ± 8%	**33** ± 8%
DDK	**25** ± 7%	29 ± 8%

TABLE 6.9 – EER (moyenne ± écart type) issus de la classification MP hommes vs sains hommes, à partir des tâches de répétitions de phrase, monologue et diadococinésie des enregistrements téléphoniques. Comparaison synthétique entre classification MFCC-GMM et x-vecteurs classés avec LDA + distance cosinus. Augmentation de données pour entraînement LDA des tâches indépendantes du texte (soit monologue).

L'ensemble des résultats est présenté au Tableau 6.10, avec un résumé au Tableau 6.11.

Comme pour les hommes, nous retrouvons une amélioration des performances avec l'apport de l'augmentation de données pour la LDA et PLDA seulement pour la tâche texte-indépendant. De même nous retrouvons de meilleures performances pour la LDA + distance cosinus et pour la PLDA, que pour la distance cosinus simple. Enfin nous retrouvons une amélioration par rapport à la méthode MFCC-GMM, qui est même plus importante que pour les hommes. Les x-vecteurs classés avec une LDA + distance cosinus améliorent en effet de 7 à 8% les performances obtenues avec la méthode MFCC-GMM, pour les tâches de répétition de phrases et monologue (cf. Tableau 6.11).

Tâche	MFCC-GMM	dist cos	LDA + dist cos	PLDA	LDA aug + dist cos	PLDA aug
repet	42 %	49 %	**34** %	34 %	39 %	37 %
monol	40%	43 %	34 %	36 %	**33** %	33 %

TABLE 6.10 – EER issus de la classification MP femmes vs sains femmes, à partir des tâches de répétitions de phrase, et monologue des enregistrements téléphoniques. Comparaison types de classifications et effet augmentation de données.

Tâche	MFCC-GMM	LDA + dist cos
repet	42 ± 10%	**34** ± 9%
monol	40 ± 10%	**33** ± 9 %

TABLE 6.11 – EER (moyenne ± écart type) issus de la classification MP femmes vs sains femmes, à partir des tâches de répétitions de phrase, et monologue des enregistrements téléphoniques. Comparaison synthétique entre classification MFCC-GMM et x-vecteurs classés avec LDA + distance cosinus. Augmentation de données pour entraînement LDA des tâches indépendantes du texte (soit monologue).

La comparaison des performances issues des différentes méthodes de classification à partir du monologue peut également être visualisée par la courbe DET Figure 6.1. Les résultats concernant la LDA et la PLDA sont ceux obtenus avec l'augmentation de données. Nous pouvons constater que les classifications avec LDA et PLDA ont de meilleures performances que la distance cosinus simple et la classification MFCC-GMM.

6.2.3 Classification avec le microphone professionnel

Enfin nous avons testé les méthodes de classification avec les x-vecteurs, en utilisant cette fois-ci les enregistrements issus du microphone professionnel. Le modèle DNN pré-entrainé que nous avons utilisé est le modèle voxeleb présenté dans la partie méthode. Pour la classification

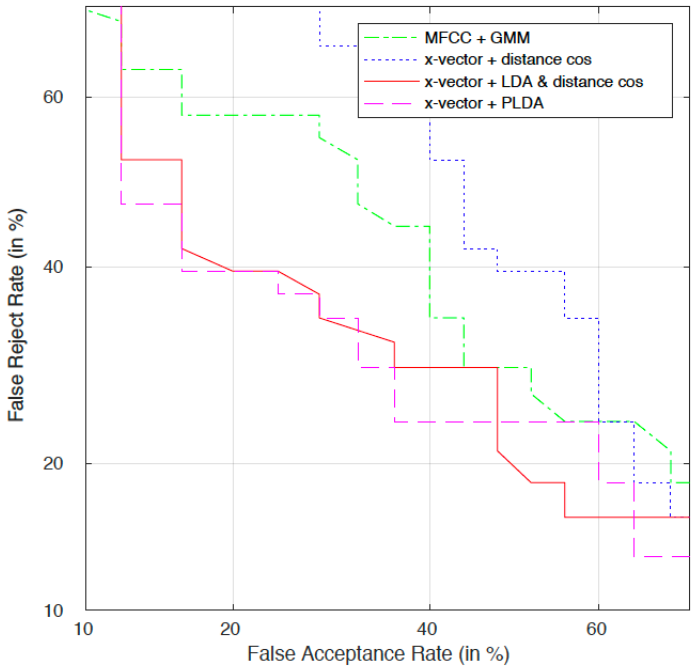


FIGURE 6.1 – Courbes DET issues de la classification femmes MP vs femmes sains à partir du monologue en enregistrement téléphonique. Comparaison entre 4 méthodes de classification : MFCC+ GMM, x vecteurs + distance cosinus, x-vecteurs + LDA et distance cosinus, et PLDA. La LDA et PLDA ont été entrainées avec une augmentation de données.

des hommes nous avons utilisé 36 MP et 36 sains à chaque run pour l'entraînement de la LDA et PLDA et la constitution des x-vecteurs MP et sain de références. Le reste des sujets hommes (38 MP et 12 sains) constituait le groupe test. Pour le monologue cela fait un total moyen de 1h10 de données voix pour le groupe d'entraînement à chaque run, et un total de 1min par sujet test. Pour les lecture et répétitions de phrases, cela revient à environ 1h20 pour le groupe d'entraînement et 1min10 par sujet test.

Pour les femmes le groupe d'entraînement comprenait 30 MP et 30 sains et le groupe test 11 MP et 13 sains. Pour le monologue cela fait un total moyen de 1h de données voix pour le groupe d'entraînement à chaque run, et un total de 1min par sujet test. Pour les lecture et répétitions de phrases, cela revient à environ 1h10 pour le groupe d'entraînement et 1min10 par sujet test.

L'ensemble des résultats est présenté Tableau 6.12 pour les hommes et au Tableau 6.14 pour les femmes. Les Tableaux 6.13 et 6.15 synthétisent ces derniers.

Nous pouvons constater une très légère amélioration de la classification par x-vecteurs avec LDA + distance cosinus par rapport à la classification MFCC-GMM, pour les hommes. Pour les femmes, cette amélioration est beaucoup plus marquée, surtout pour le monologue pour lequel on observe un gain de performance de 15%.

Tâche	MFCC-GMM	cos	LDA + cos	PLDA	LDA aug + cos	PLDA aug
repet & lect	22 %	32 %	**22** %	24 %	24 %	25 %
monol	26%	35 %	27 %	28 %	**25** %	25 %

TABLE 6.12 – EER issus de la classification hommes MP vs sain, à partir des tâches de répétitions et lecture de phrases, et du monologue des enregistrements issus du microphone professionnel. Comparaison types de classifications et effet augmentation de données.

Tâche	MFCC-GMM	LDA + dist cos
repet & lecture	22 ± 6%	**22** ± 6%
monologue	26 ± 6%	**25** ± 6%

TABLE 6.13 – EER (moyenne ± écart type) issus de la classification des hommes MP vs sain, à partir des tâches de répétitions et lecture de phrases, et du monologue des enregistrements issus du microphone professionnel. Comparaison synthétique entre classification MFCC-GMM et x-vecteurs classés avec LDA + distance cosinus. Augmentation de données pour entraînement LDA des tâches indépendantes du texte (soit monologue).

Tâche	MFCC-GMM	cos	LDA + cos	PLDA	LDA aug + cos	PLDA aug
repet & lect	42 %	51 %	39 %	39 %	34 %	33 %
monol	45%	41 %	32 %	35 %	**30** %	30 %

TABLE 6.14 – EER issus de la classification des femmes MP vs sain, à partir des tâches de répétitions et lecture de phrases, et du monologue des enregistrements issus du microphone professionnel. Comparaison types de classifications et effet augmentation de données.

Tâche	MFCC-GMM	LDA + dist cos
repet & lecture	42 ± 8%	**39** ± 7%
monologue	45 ± 8%	**30** ± 7%

TABLE 6.15 – EER (moyenne ± écart type) issus de la classification des femmes MP vs sain, à partir des tâches de répétitions et lecture de phrases, et du monologue des enregistrements issus du microphone professionnel. Comparaison synthétique entre classification MFCC-GMM et x-vecteurs classés avec LDA + distance cosinus. Augmentation de données pour entraînement LDA des tâches indépendantes du texte (soit monologue).

La comparaison des performances issues des différentes méthodes de classification à partir du monologue enregistré avec le microphone professionnel peut également être visualisée par la courbe DET Figure 6.2. Les résultats concernant la LDA et la PLDA sont ceux obtenus avec l'augmentation de données. Nous pouvons constater, comme pour les enregistrements téléphoniques, que les classifications avec LDA et PLDA ont de meilleures performances que la distance cosinus simple et la classification MFCC-GMM.

6.3 Conclusion x-vecteurs

Nous avons adapté la dernière méthode en date utilisée en reconnaissance du locuteur, en encore jamais utilisée dans le cadre de la détection de MP. Cette méthode se base sur l'extraction d'embeddings, appelés x-vecteurs, extraits à partir d'un DNN prenant en entrée des vecteurs MFCC. Nous avons fait varier différentes conditions, tout en comparant, pour chaque condition, 3 méthodes de classification (distance cosinus, LDA + distance cosinus et PLDA). Comme l'entraînement du DNN nécessite généralement beaucoup de données, nous avons utilisé un DNN pré-entrainé pour la reconnaissance du locuteur.

Les analyses sur notre base téléphonique concernant la classification des hommes MP vs sains MP, nous ont permis de constater que les performances étaient meilleures quand les segments audio testés avaient la même durée (3s) que les segments ayant servi pour l'entraînement (du DNN, de la LDA et de la PLDA) et pour la constitution des x-vecteurs moyens MP et sain. Même si isolement, les segments de courtes durées sont en général moins bien classés que ceux de longue durée [Snyder et al., 2017, Snyder et al., 2018a], le score moyen se fait à partir de plus

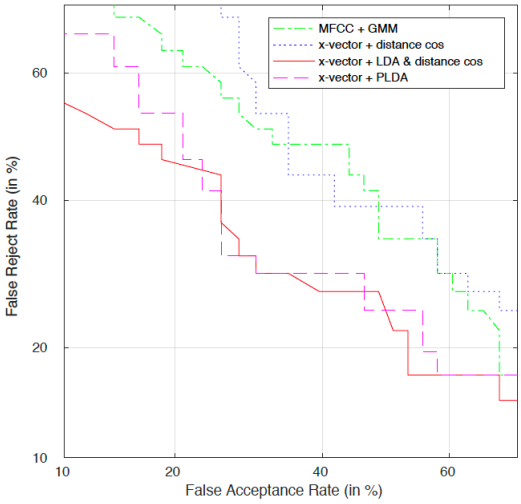


FIGURE 6.2 – Courbes DET issues de la classification des femmes MP vs sain à partir du monologue enregistré avec le microphone professionnel. Comparaison entre 4 méthodes de classification : MFCC + GMM, x vecteurs + distance cosinus, x-vecteurs + LDA et distance cosinus, et PLDA.

de tests quand les segments sont plus courts, permettant ainsi d'améliorer les résultats pour des segments tests d'une durée de quelques secondes, comparable à celle des segments d'entraînement.

Concernant la comparaison des différents types de classifications, les observations sur l'ensemble des conditions d'enregistrements (téléphone ou microphone professionnel) et pour les deux sexes (hommes et femmes) montrent dans l'ensemble une nette amélioration des performances quand on ajoute une LDA avant le calcul de distance cosinus. On constate également une performance équivalente entre LDA + cos distance et la PLDA.

L'ajout de l'augmentation de données améliore les performances du monologue pour toutes conditions d'enregistrements et les groupes. Les performances des tâches texte dépendantes telles que les répétitions de phrases et lecture ne sont dans l'ensemble pas améliorées avec l'augmentation de données, ce qui est cohérent avec le fait que l'augmentation de données en rajoutant du bruit de différentes sortes, nuit à la spécificité du contenu phonétique.

Nous avons effectué nos analyses sur 3 types de tâches : le monologue, la répétition de phrase (et lecture), et les tâches DDK et avons comparé les résultats avec les performances obtenues avec l'analyse MFCC-GMM effectuée précédemment. Nous avons constaté, pour toutes les conditions d'enregistrements et tous les groupes, une amélioration des performances de classification pour la tâche de monologue (texte-indépendant), cf. Tableau 6.16. Ce qui est cohérent avec le fait que les x-vecteurs ont été à l'origine élaborés pour la reconnaissance du locuteur indépendante du texte. L'amélioration sur les tâches dépendantes du texte (répétition de phrase et lecture) apparaît également mais de manière moins prononcée dans nos analyses, cf. Tableau 6.17. Enfin les tâches très spécifiques, comme les DDK, présentent de meilleures performances avec les GMM qu'avec les x-vecteurs. Ceci pouvant être la conséquence du DNN pré-entrainé pour la reconnaissance du locuteur à partir de données paroles beaucoup plus variées que les phonèmes prononcés lors les tâches DDK.

Dans le but de rendre le DNN plus spécifique aux tâches DDK, nous avons effectué une ana-

	classes	MFCC-GMM	x-vecteur (LDA + dist cos)
téléphone	MPh vs sainh	36 ± 8%	33 ± 8%
	MPf vs sainf	40 ± 10%	33 ± 9%
micro pro	MPh vs sainh	26 ± 6%	25 ± 6%
	MPf vs sainf	45 ± 8%	30 ± 7%

TABLE 6.16 – Comparaison des EER (moyenne ± écart type) obtenus avec la méthode MFCC-GMM et avec la méthode x-vecteurs classés avec LDA + distance cosinus, pour la détection de MP chez les hommes (MPh vs sainh) et chez les femmes (MPf vs sainf). La tâche utilisée est le **monologue**. Une augmentation de données est effectuée pour la méthode x-vecteurs.

	classes	MFCC-GMM	x-vecteur (LDA + dist cos)
téléphone	MPh vs sainh	35 ± 8%	32 ± 8%
	MPf vs sainf	42 ± 10%	34 ± 9%
micro pro	MPh vs sainh	22 ± 6%	22 ± 6%
	MPf vs sainf	42 ± 8%	39 ± 7%

TABLE 6.17 – Comparaison des EER (moyenne ± écart type) obtenus avec la méthode MFCC-GMM et avec la méthode x-vecteurs classés avec LDA + distance cosinus, pour la détection de MP chez les hommes (MPh vs sainh) et chez les femmes (MPf vs sainf). La tâche utilisée est la **répétition de phrases** (avec lecture) pour les enregistrements du microphone professionnel.

lyse en l'entrainant avec notre base de données (en utilisant les tâches DDK). Les performances obtenues n'ont pas montré d'amélioration par rapport au DNN pré-entrainé pour la reconnaissance du locuteur. Ceci pouvant être dû à la quantité réduite de nos données disponibles pour l'entraînement du DNN (nécessitant habituellement beaucoup de données).

Enfin le dernier point à souligner est la nette amélioration des performances, par rapport à la méthode MFCC-GMM, avec la méthode x-vecteurs + LDA, pour la détection de MP chez les femmes à partir du monologue (cf. Tableau 6.16). On trouve une amélioration de l'EER de l'ordre de 10% (7% pour les enregistrements téléphoniques et 15% pour le microphone professionnel). Cette amélioration pourrait provenir de la LDA qui diminue la variabilité intraclasse, connue pour être importante avec les paramètres types MFCC chez les femmes [Fraile et al., 2009b].

Ces deux types de classification inspirée de la reconnaissance du locuteur (MFCC-GMM et x-vecteur) permettent une détection de la maladie de Parkinson au stade débutant en exploitant quasiment uniquement les troubles articulatoires. Or les altérations vocales rencontrées dans la maladie de Parkinson ne concernent pas seulement l'articulation, mais aussi la prosodie, la phonation, le débit de parole et les habilités rythmiques. Nous avons donc voulu analyser également ces autres domaines afin d'enrichir les informations vocales dont nous pouvons disposer pour détecter MP précocement.

Chapitre 7

Classification MP vs sain à partir de paramètres globaux

Dans ce chapitre nous présenterons dans un premier temps les paramètres globaux que nous avons extraits et les méthodes d'extraction utilisées. Ensuite nous étudierons si ces paramètres diffèrent de manière significative entre les groupes MP et sain, en effectuant des analyses de variance. Enfin nous détaillerons la méthode de classification utilisée à partir de ces paramètres et les performances obtenues.

7.1 Extraction des paramètres

Différents paramètres vocaux ont été extraits des différentes tâches, à l'aide des logiciels Praat [Boersma and Weenink, 2001] et Matlab (MathWorks Inc, Natick, MA). Ces paramètres décrivent la phonation, la prosodie, la répartition des pauses et la capacité à suivre un rythme. Comme pour les analyses précédentes, une étape de prétraitement par soustraction spectrale a été effectuée avant l'extraction des paramètres pour débruiter le signal.

7.1.1 Prosodie

Les variations de hauteur et d'intensité de la voix constituent ce qu'on appelle l'intonation ou prosodie, et permettent de transmettre des émotions, des intentions et apportent un sens complémentaire aux mots.

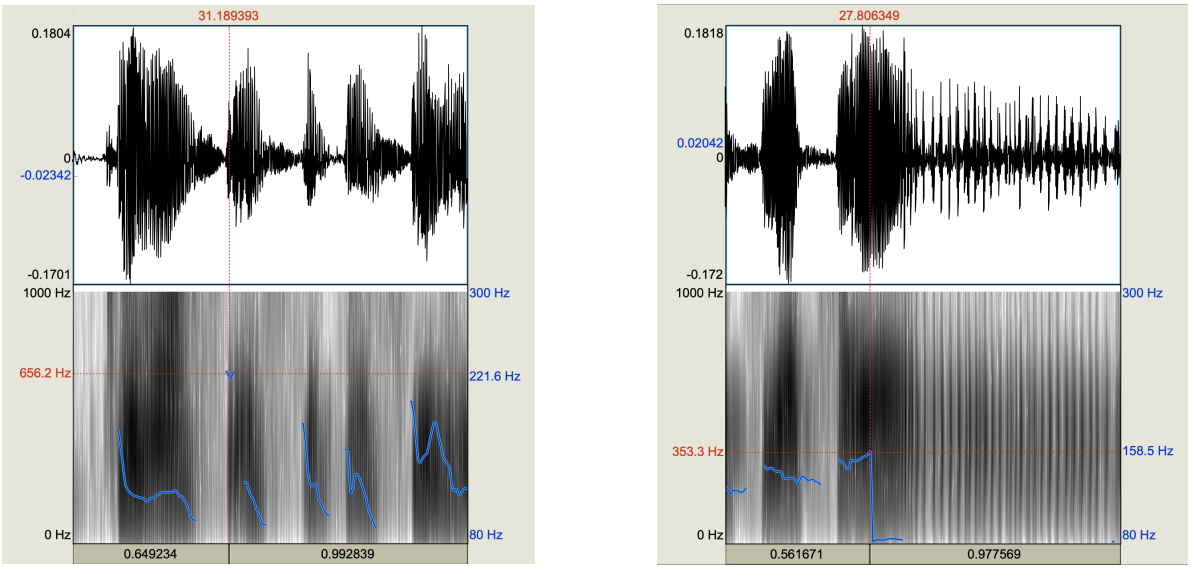
7.1.1.1 Variation de la hauteur de la voix

La variation de la hauteur de la voix, responsable de la mélodie, est quantifiée ici par l'écart type de la fréquence fondamentale (SD log Fo), extraite pendant les tâches de lecture, répétition de phrases et de monologue. Nous considérons plus précisément l'écart type du logarithme de Fo et non directement de Fo pour que l'écart type dépende moins de la hauteur moyenne de la voix, prenant en compte le fait que l'étalement de la plage de fréquences de Fo suit une loi exponentielle selon le Fo moyen.

Une méthode d'auto-corrélation, telle que décrite dans [Boersma, 1993] a été utilisée pour extraire Fo, cette méthode est particulièrement adaptée pour l'extraction de Fo dans le cadre de l'analyse de la prosodie. L'absence ou la présence de son voisé est estimée environ toutes les 10ms avec l'estimation de Fo lorsque la trame est considérée comme voisée. Cependant l'extraction n'est pas parfaite et peut contenir des valeurs erronées, comme des Fo aigus alors que le son n'est pas voisé, et des *halving* (Fo/2 à la place de Fo). Ces valeurs aberrantes se rencontrent d'autant plus pour les voix rauques et cassées, et faussent (en l'augmentant) la valeur de SD log

Fo. Afin de supprimer un maximum de ces valeurs aberrantes, nous avons restreint la fenêtre fréquentielle d'extraction à [80-300Hz] pour les hommes et à [100-600Hz] pour les femmes, et avons choisi un seuil de voisement élevé.

Nous avons ensuite créé un algorithme de suppression de valeurs aberrantes qui :
— supprime les Fo isolés (1 ou 2 Fo entourés de sons non voisés) ;
— supprime les Fo aigus anormaux (écart important par rapport au Fo précédent, et par rapport à la moyenne, ces Fo sont généralement la conséquence de bruits de bouche ou de gorge) cf. Figure 7.1a ;
— supprime les *halving*, par le biais d'un filtre passe haut dépendant de la moyenne (les halving sont souvent rencontrés quand la voix est craquelée) cf. Figure 7.1b.



(a) Fo aigu erroné (b) halving

FIGURE 7.1 – Tracé du signal sonore avec spectrogramme (échelle des fréquences à gauche) et estimation de Fo en bleu (échelle des fréquences à droite). Cas de deux patients MP avec la voix rauque. A gauche problème de Fo aigu erroné, à droite problème de halving lors d'un "euh" craquelé. Ces deux erreurs de Fo se trouvent au niveau des curseurs rouges et sont supprimées postérieurement par l'algorithme de suppression des valeurs aberrantes.

Ce postprocessing supprime quasiment la totalité des erreurs d'estimation de Fo pour les voix saines. Pour les voix très éraillées (comme on peut rencontrer chez les MP à un stade un peu avancé) l'estimation de Fo est moins fiable, et peut entraîner des valeurs élevées de SD log Fo qui ne correspondent pas à la réelle prosodie. Néanmoins nous avons choisi de garder tous les sujets analysés précédemment pour cette analyse pour être sûrs de ne pas induire de biais optimiste dans les résultats.

7.1.1.2 Variations d'intensité

Les variations d'intensités, servant notamment à mettre l'accent sur certains mots, sont quantifiées par ici par l'écart type de l'intensité (*SD Int*). Bien que, comme pour Fo, les modulations d'intensités de la voix suivent une loi exponentielle suivant la valeur moyenne de l'intensité, il n'a pas été nécessaire de prendre l'écart type du logarithme, car l'intensité est exprimée en dB, qui suit déjà une échelle logarithmique.

Nous avons extrait l'intensité avec le logiciel Praat toutes les 8ms. L'intensité est calculée comme le carré de la pression acoustique convoluée avec une fenêtre temporelle gaussienne de

32ms. Une fois les intensités extraites, nous avons gardé seulement celles correspondant à des trames sonores. Ceci de manière à avoir les modulations d'intensité lors de la parole, sans que la fréquence ni la durée des pauses ne modifient *SD Int.*

7.1.2 Phonation

Afin de caractériser le timbre de la voix des sujets MP et des sujets sains, nous avons extrait différents paramètres liés à la phonation, pendant la tâche de voyelle soutenue /a/ :

- La fraction de trames localement non voisées (DUV pour *Degree of Unvoiced frames*) et la durée des segments non voisés divisée par la durée totale (DVB pour *Degree of Voice Breaks*) caractérisent les cassures dans la voix. Nous avons extrait ces paramètres sur la portion de la tâche où le sujet prononce la voyelle soutenue, afin de ne pas prendre en compte les silences de début et de fin de tâche, qui augmenteraient artificiellement ces deux paramètres.

- Les jitters caractérisent les fluctuations courts termes de la fréquence fondamentale. Le jitter local est défini comme la moyenne de toutes les différences entre les durées de deux périodes consécutives du signal divisée par la durée moyenne d'une période du signal. On considère généralement 1.04% comme le seuil pathologique. Le jitter RAP (*Relative Average Perturbation*) se calcule en prenant la moyenne des différences entre une période Ti et la moyenne des 3 périodes successives Ti-1, Ti et Ti+1, toujours divisée par la période moyenne. Ceci a pour effet d'atténuer les variations volontaires de la fréquence vocale, comme le trémolo, par exemple. Le jitter PPQ5 (5-point Period Perturbation Quotient) se calcule de la même manière que le RAP mais en prenant la moyenne de 5 périodes consécutives au lieu de 3. Le dernier jitter est le jitter DDP (Différence des Différences de Période). Il s'exprime comme la différence entre deux différences consécutives de deux périodes consécutives, divisée par la période moyenne.

- Les shimmers caractérisent les fluctuations courts termes de l'intensité sur le même principe que les jitters pour la fréquence. Le shimmer local est la différence entre deux amplitudes de périodes consécutives divisée par l'amplitude moyenne. Le seuil pathologique est fixé à 3.81%. Le shimmer APQ3 (*3-point Amplitude Perturbation Quotient*) est la différence entre l'amplitude d'une période et la moyenne de l'amplitude de cette période et des deux périodes qui l'entourent, divisée par l'amplitude moyenne. Les shimmers APQ5 et APQ11 fonctionnent de la même manière mais en prenant respectivement la moyenne de 5 et 11 amplitudes consécutives au lieu de 3. Le shimmer DDP est la différence entre deux différences consécutives d'amplitudes consécutives, divisée par l'amplitude moyenne.

- Les ratios HNR (*Harmonics-to-Noise Ratio*) et NHR (*Noise-to-Harmonics Ratio*) caractérisent l'harmonicité de la voix, c'est à dire la proportion de l'énergie contenue dans les harmoniques par rapport à celle contenue dans le bruit (le reste du spectre). Le ratio HNR sera faible (et NHR grand) pour les voix qu'on qualifierait d'éraillées ou soufflées.

Ces paramètres de phonation dépendent tous de l'estimation de la fréquence fondamentale. Cette dernière a été calculée par une méthode de corrélation croisée sur une étendue de 75 à 300Hz pour les hommes et de 100 à 400Hz pour les femmes.

7.1.3 Pauses

Le débit de parole est connu pour être altéré dans la maladie de Parkinson avec une impression à l'oreille de débit saccadé. De manière à quantifier cette altération nous avons étudié le nombre et la durée des pauses lors du monologue et de la lecture chez les sujets MP, sains et

iRBD. Nous nous sommes intéressés aux pauses silencieuses (avec ou sans respiration), et non aux pauses sonores (procédé d'hésitation ("euh", "mm"..), allongement des syllabes, répétitions..) car plus difficiles à détecter de façon automatique.

La première étape a consisté à détecter les trames silencieuses. Nous sommes partis d'un signal filtré entre 80 et 8000Hz, à partir duquel l'intensité à été calculée environ toutes les 10ms. Le filtrage passe bande sert à limiter l'effet des bruits sur l'estimation de l'intensité. Nous avons ensuite considéré comme silencieuse toute trame dont l'intensité était inférieure à l'intensité maximale moins un certain seuil. Le choix de ce seuil est délicat, car un seuil trop faible entraine une surdétection de silences, notamment les consonnes à faibles intensité, comme /s/ et /ch/ sont considérées comme des silences. Un seuil trop élevé entraine quant à lui, une sous détection de pauses, notamment les bruits de respiration sont alors considérés comme des sons. Le seuil parfait, avec zéro erreur n'existe pas mais le seuil de 25dB s'est révélé être le meilleur compromis.

Nous avons ensuite calculé les durées des pauses, et les avons regroupées en plusieurs tranches : de 200 à 500ms, de 500ms à 1s et supérieures à 1s. Nous avons aussi comptabilisé le nombre de pauses de plus de 200ms et leur durée médiane. Nous avons préféré la médiane à la moyenne, pour éviter que quelques pauses trop longues (dues à un manque d'inspiration par exemple) ne viennent fausser le résultat. Nous n'avons considéré que les pauses supérieures à 200ms car une étude de quelques sujets nous ont permis de constater que la détection des pauses inférieures à 200ms n'était pas assez fiable, détectant trop de silences lors de la prononciation de consonnes. Enfin de manière à ce que le nombre de pauses soit indépendant de la durée parlée, nous l'avons divisé par cette dernière.

7.1.4 Rythme

Comme vu partie 2.3.4, une autre perturbation liée à la maladie de Parkinson détectable relativement tôt par des exercices vocaux concerne les capacités à répéter des sons avec un rythme régulier. Nous avons voulu tester si nos patients rencontraient ces difficultés, et si ces difficultés étaient accrues suivant le type de tâches.

Pour cela nous avons analysé les tâches de répétions lentes de syllabes /pa/, /kou/ et /pa kou/. Pour rappel les sujets entendent les syllabes /pa/ à un rythme régulier d' une syllabe par seconde, et ont comme consigne de répéter ces syllabes sur le même rythme, une fois l'exemple terminé, pendant 30s. La même tâche est faite avec les syllabes /kou/, puis en alternant les syllabes /pa/ et /kou/. Cette dernière tâche est faite deux fois avec le microphone professionnel (et donc le microphone de l'ordinateur) et une fois par session téléphonique.

Nous avons extrait, à partir de ces tâches, un paramètre décrivant la capacité à maintenir un rythme constant et un autre paramètre caractérisant la capacité à suivre un rythme imposé. Nous avons commencé par effectuer une VAD, en utilisant l'extraction des Fo du logiciel praat. Les estimations de Fo n'ayant pas nécessité d'être très précises, car servant juste à détecter les trames voisées, nous avons utilisé le seuil de voisement par défaut (moins élevé que celui utilisé pour l'analyse de SD log Fo) et nous avons utilisé des plages de fréquences plus larges, vu que les éventuels halving n'impactaient pas l'analyse ici. Nous avons ensuite utilisé le même algorithme que pour l'extraction des transitions non voisé à voisé, présenté partie 5.4, pour connaitre les temps de début de chaque syllabe. Afin de limiter les risques de fausses détections de débuts de syllabes, nous avons au préalable effectué un prétraitement pour supprimer les valeurs isolées de Fo, en ne gardant que les paquets d'au minimum 6 trames voisées consécutives. En effet, certaines attaques de syllabes, notamment la syllabe /kou/, peuvent contenir quelques trames identifiées comme non voisées, donc avec une absence de Fo, risquant d'entraîner une détection supplémentaire erronée de début de syllabes. Avec cette méthode de prétraitement et

de détection d'attaque de sons voisés, nous limitons ces anomalies. Néanmoins il peut arriver que pour certains rares sujets avec la voix rauque, certaines occurrences de syllabes ne soient pas détectées, impactant l'estimation des paramètres de variations du rythme pour la tâche en question. Ces altérations sont néanmoins atténuées quand les paramètres sont moyennés sur plusieurs tâches.

Une fois les temps de débuts de syllabes extraits, nous avons pu calculer les durées d_i entre chaque syllabe et ainsi évaluer le rythme auquel elles sont prononcées. Nous avons calculé, pour chaque sujet, l'écart type relatif (RSD pour Relative Standard Deviation), défini par l'équation :

$$RSD = \frac{SD}{\mu} \tag{7.1}$$

Avec SD l'écart type des durées d_i et μ la moyenne des d_i. RSD est une mesure de dispersion relative, qui permet d'évaluer la capacité d'un sujet à maintenir un rythme constant (quelle que soit sa valeur moyenne) pendant l'exécution des tâches de répétitions lentes de syllabes.

Le deuxième paramètre (Ecart) que nous avons extrait est l'écart (en valeur absolue) entre la durée moyenne des d_i d'un sujet et la durée moyenne d_{ex} de l'exemple à suivre. Ce paramètre caractérise la capacité du sujet à suivre globalement le rythme imposé.

7.2 Analyses de Variance

Afin de savoir quels paramètres parmi ceux extraits sont discriminants dans la maladie de Parkinson, nous effectué plusieurs analyses de variance. Pour comparer les groupes MP et sain nous avons utilisé le test statistique dit Welch t-test, qui permet de comparer deux moyennes provenant de deux échantillons distincts (sujets différents) de variances inégales. Nous considérons l'hypothèse nulle d'égalité des moyennes rejetées si la valeur p obtenue est inférieure à un certain seuil de significativité. Ce seuil est généralement fixé à 5% ce qui correspond un risque de 5% de rejeter à tort l'hypothèse nulle pour une variable en particulier. Dans le cas de comparaisons multiples, la probabilité d'avoir une variable pour laquelle l'hypothèse nulle est rejetée à tort augmente avec le nombre de variables testées. Il convient alors d'effectuer une compensation en abaissant le seuil de significativité. Nous avons ainsi considéré 0.001 comme seuil de significativité, à la place de 0.05. Enfin nous avons séparé les hommes des femmes pour les analyses de manière à mieux identifier les paramètres discriminants pour chaque genre.

7.2.1 Prosodie

7.2.1.1 Variations de la hauteur de la voix

Voici les résultats concernant les variations de la hauteur de la voix, caractéristiques de la mélodie, quantifiées par le paramètre SD (log Fo). Nous avons analysé ce paramètre pour différents types de tâches vocales, et avec différents types de microphones. Nous avons comparé les résultats pour les groupes MP, sains et iRBD, en séparant toujours les hommes des femmes, afin d'étudier l'effet par genre.

Comparaison des tâches Les moyennes et écart types du paramètre prosodique SD log Fo pour les groupes MP et sains enregistrés avec le microphone professionnel sont détaillées Figure 7.2. SD log Fo a été successivement extrait pour les tâches de lecture, répétitions de phrases et pour le monologue. La dernière ligne du tableau correspond à la moyenne des valeurs obtenues pour la lecture et les répétitions de phrases avec les valeurs obtenues lors du monologue. Nous constatons que pour toutes ces tâches, la valeur de prosodie diminue chez les MP, comparé

aux sains, aussi bien chez les hommes que chez les femmes, mais de façon plus prononcée chez les hommes. Cette diminution traduit l'aspect monotone de la voix que l'on rencontre dans la maladie de Parkinson.

Chez les hommes la diminution est significative pour toutes les tâches. Le contenu émotionnel (caractérisé par les nombreuses phrases interrogatives et exclamatives du dialogue et des tâches de répétitions) semble augmenter les différences prosodiques entre les MP et les sains, par rapport au contenu neutre (lecture de texte), ce qui va dans les sens de [Rusz et al., 2011a].

	Hommes					Femmes				
	MP		sain		valeur-p	MP		sain		valeur-p
	moy	SD	moy	SD		moy	SD	moy	SD	
lecture	0.0624	0.0134	0.0766	0.0164	**7.6E-07**	0.0734	0.0149	0.0795	0.0169	0.084
dont dialogue	0.0591	0.0153	0.0748	0.0158	**2.6E-07**	0.0661	0.0152	0.0747	0.0154	**0.012**
dont texte	0.0544	0.0136	0.0657	0.0154	**4.1E-05**	0.0597	0.0188	0.0662	0.0142	0.081
répétition	0.0686	0.0124	0.0812	0.0111	**7.9E-08**	0.0813	0.0140	0.0877	0.0147	**0.045**
monologue	0.0501	0.0111	0.0662	0.0172	**5.6E-09**	0.0644	0.0166	0.0693	0.0140	0.15
moy	0.0578	0.0101	0.0725	0.0128	**1.0E-10**	0.0709	0.0136	0.0764	0.0119	**0.049**

FIGURE 7.2 – Moyennes et écarts types du paramètre prosodique SD(log Fo), et valeur-p issues du Welch t-test. Influence des différents types de tâches vocales. La dernière ligne correspond à la moyenne des valeurs obtenues pour la lecture et les répétitions de phrases avec les valeurs obtenues lors du monologue.

Comparaison des microphones Nous avons comparé les performances obtenues avec les différents microphone (microphone professionnel, microphone de l'ordinateur et téléphone) en considérant le score moyen prosodique (moyenne des lectures, répétitions de phrases et monologue). Les enregistrements téléphoniques ne contenant pas de lecture, la moyenne est effectuée sur les répétitions de phrases et le monologue. De plus nous avons considéré toutes les sessions de notre base de données. Pour les sujets qui ont fait plusieurs sessions téléphoniques, le paramètre prosodique est extrait pour chaque session puis moyenné. Pour le microphone professionnel et celui de l'ordinateur nous avons en plus comparé les scores de la lecture du dialogue (tâche non effectuée au téléphone). Les résultats sont présentés Figure 7.3.

Nous constatons que les différences prosodiques entre MP et sains sont encore mieux décelées avec le microphone interne de l'ordinateur qu'avec le microphone professionnel. Ceci peut être dû à la distance plus grande entre le microphone et la bouche du locuteur et à la fonction de réduction de bruit active de l'ordinateur. Nous avons vu avec l'analyse MFCC-GMM que ces deux caractéristiques semblaient atténuer les différences spectrales lors notamment de la prononciation des occlusives, rendant la détection de MP à partir des tâches DDK plus difficile, cf. partie 5.2.2. Ces conditions d'enregistrements avec le microphone de l'ordinateur semblent par contre être bénéfiques pour l'estimation de la prosodie. En effet la distance plus importante et le débruitage actif peuvent diminuer les sons non voisés qui peuvent à tort être considérés comme voisés et entraîner des erreurs de Fo.

Concernant les enregistrements téléphoniques, la comparaison est plus délicate car le nombre de sujets et la quantité de données voix par sujet et les tâches vocales sont légèrement différents. Le score que nous avons considéré est la moyenne des scores prosodiques des tâches de répétitions de phrases et du monologue. Nous constatons toujours une diminution de la prosodie pour les sujets MP mais les différences entre les valeurs p sont difficilement interprétables. Néanmoins on peut constater que la diminution de SD log Fo pour les MP reste significative chez les hommes avec ces enregistrements téléphoniques.

		Hommes					Femmes				
Micro	Tâches	MP		sain		valeur-p	MP		sain		valeur-p
		moy	SD	moy	SD		moy	SD	moy	SD	
micro pro	dialogue	5,91E-02	1,53E-02	7,48E-02	1,58E-02	**2.6E-07**	6,61E-02	1,52E-02	7,47E-02	1,54E-02	**0.012**
	moy	5,78E-02	1,01E-02	7,25E-02	1,28E-02	**1.0E-10**	7,09E-02	1,36E-02	7,64E-02	1,19E-02	**0.049**
micro ordi	dialogue	5,57E-02	1,60E-02	7,37E-02	1,59E-02	**1,50E-08**	6,23E-02	1,51E-02	7,25E-02	1,44E-02	**2,1E-03**
	moy	5,45E-02	1,10E-02	7,20E-02	1,54E-02	**3,2E-11**	6,76E-02	1,48E-02	7,41E-02	1,29E-02	**0.035**
téléphone	moy	6,14E-02	1,13E-02	7,57E-02	1,14E-02	**2,6E-08**	7,64E-02	1,24E-02	8,51E-02	1,22E-02	**7,7E-03**

FIGURE 7.3 – Moyennes et écarts types du paramètre prosodique SD(log Fo), et valeur-p issues du Welch t-test. Influence des différents types de microphones. La moyenne prend en compte les tâches de répétitions de phrases et monologue pour les 3 types de microphones ainsi que les tâches de lecture (tâche non faite au téléphone) pour les microphones professionnel et ordinateur.

Cas des iRBD Concernant les iRBD hommes les résultats du paramètre SD (log Fo), calculés sur les différentes tâches enregistrées avec le microphone professionnel, sont présentés Figure 7.4. Pour la comparaison nous avons fait refigurer les résultats des hommes MP et sain. N'ayant que 4 iRBD femmes dans notre base de données nous n'avons pas fait d'analyse avec car elle n'aurait pas eu de valeur statistique. Les boites à moustaches représentées Figure 7.5 illustrent les répartitions des valeurs de SD (log Fo) au sein des groupes hommes MP, iRBD et sains.

Tâches	MP		iRBD		sain		p (iRBD vs MP)	p(iRBD vs sain)
	moy	SD	moy	SD	moy	SD		
lecture	6,24E-02	1,34E-02	7,15E-02	1,73E-02	7,66E-02	1,64E-02	**2,3E-03**	0.16
dont dialogue	5,91E-02	1,53E-02	6,92E-02	1,86E-02	7,48E-02	1,58E-02	**2,1E-03**	0.13
dont texte	5,44E-02	1,36E-02	6,40E-02	1,49E-02	6,57E-02	1,54E-02	**6,9E-04**	0.59
répétition	6,86E-02	1,24E-02	7,78E-02	1,55E-02	8,12E-02	1,11E-02	**7,3E-04**	0,23
monologue	5,01E-02	1,11E-02	6,34E-02	1,80E-02	6,62E-02	1,72E-02	**3,1E-06**	0,46
moy	5,78E-02	1,01E-02	6,90E-02	1,47E-02	7,25E-02	1,28E-02	**4,2E-06**	0,23

FIGURE 7.4 – Moyennes et écarts types du paramètre prosodique SD(log Fo) extraits sur différentes tâches avec le microphone professionnel. Les valeur-p sont issues du Welch t-test comparant les groupes hommes iRBD et MP d'une part et les groupes hommes iRBD et sain d'autre part. La moyenne prend en compte les tâches de lecture, répétitions de phrases et le monologue.

Nous constatons d'après la Figure 7.4 que pour toutes les tâches le paramètre prosodique du groupe iRBD se situe entre celui du groupe MP et du groupe sain, la différence étant significative avec le groupe MP. Vu que les iRBD peuvent être considérés au stade prodromique de la maladie de Parkinson, ce résultat va dans le sens d'une diminution progressive de la prosodie au début de la maladie, qui commencerait dès le stade prodromique.

7.2.1.2 Variations de l'intensité

Nous avons calculé l'écart type de l'intensité (*SD Int*) à partir de la lecture du dialogue et du texte, et à partir du monologue, enregistrés avec le microphone professionnel. Les résultats sont présentés à la Figure 7.6.

Contrairement à ce à quoi on aurait pu s'attendre, nous ne constatons pas de diminution significative des variations d'intensité pour le groupe MP. La tâche de la lecture du dialogue, conçue notamment pour tester la capacité d'accentuation, par la mise en majuscule de certains mots, montre une légère diminution de *SD Int* chez les MP mais non significative.

Même si une diminution des variations d'intensité a été rencontrées chez les MP dans cer-

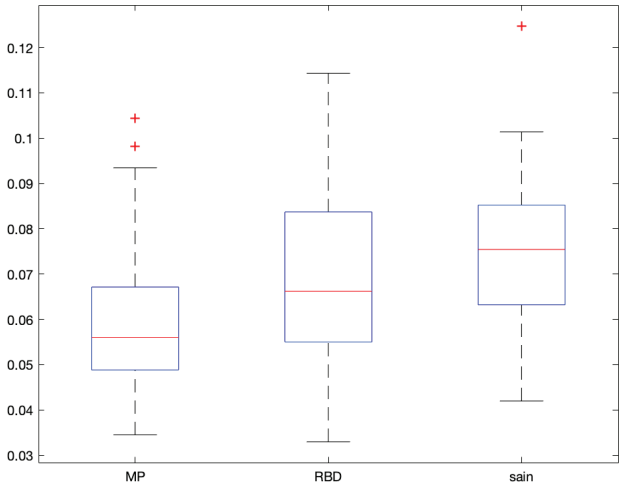


FIGURE 7.5 – Boites à moustaches représentant la répartition des valeurs de SD (log Fo) extraites lors de la lecture du dialogue par hommes MP, iRBD et sains enregistrés avec le microphone professionnel. Les traits rouges représentent les médianes, les rectangles bleus les percentiles 25 et 75, et les croix rouges les outliers.

	Hommes					Femmes				
Tâches	MP		sain		valeur-p	MP		sain		valeur-p
	moy	SD	moy	SD		moy	SD	moy	SD	
lect-dialogue	5,403	0,467	5,447	0,408	0,59	5,310	0,355	5,330	0,353	0,8
lect-texte	5,396	0,497	5,361	0,410	0,68	5,348	0,435	5,382	0,424	0,71
monologue	5,092	0,551	5,074	0,470	0,85	4,846	0,606	4,815	0,514	0,8

FIGURE 7.6 – Moyennes et écarts types de *SD Int*, lors de la lecture du dialogue et du texte et lors du monologue, enregistrés avec le microphone professionnel. Les valeur-p sont issues du Welch t-test.

taines études [Rusz et al., 2011a], il a aussi été montré que ce paramètre fait partie de ceux qui connaissent le plus d'améliorations lors de l'utilisation de traitements médicamenteux [Rusz et al., 2013b]. Or les sujets MP de notre base de données sont sous traitement médicamenteux et ont été enregistrés en ON, ce qui a probablement limité la diminution des modulations d'intensité.

Nous avons également extrait le paramètre SD Int à partir de la lecture du dialogue faite par les iRBD et avons obtenu une moyenne de 5.37. SD Int est donc aussi légèrement plus faible chez les iRBD que chez les sains, mais l'écart n'est pas non plus significatif ($p = 0.34$).

Le paramètre *SD Int*, tel qu'on l'a calculé, ne semble donc pas très adapté à la détection de MP au sein de notre base de données.

7.2.2 Phonation

Concernant l'analyse de la phonation, nous avons extraits les paramètres présentés partie 7.1.2 à partir des tâches de voyelles soutenues enregistrées avec le microphone professionnel. La tâche de voyelle soutenue /a/ étant effectuée deux fois, nous avons extrait puis moyenné ces paramètres sur les deux occurrences. Les résultats obtenus sont présentés Figure 7.7.

Nous pouvons constater qu'aucune différence significative n'apparait entre les groupes MP et sain. Néanmoins les paramètres caractérisant l'éraillement du timbre (DUV, DVB, $\frac{1}{HNR}$, et

	Hommes					Femmes				
Paramètres	MP		sain		valeur-p	MP		sain		valeur-p
	moy	SD	moy	SD		moy	SD	moy	SD	
DUV	0,010	0,045	0,005	0,013	0,47	0,014	0,039	0,016	0,048	0,88
DVB	0,012	0,059	0,005	0,018	0,44	0,022	0,069	0,025	0,080	0,85
jitter local	0,010	0,012	0,008	0,005	0,43	0,009	0,012	0,008	0,010	0,56
jitter rap	0,006	0,007	0,005	0,003	0,44	0,005	0,007	0,005	0,006	0,61
jitter ppq5	0,005	0,006	0,004	0,002	0,38	0,005	0,006	0,004	0,005	0,60
jitter ddp	0,017	0,022	0,014	0,010	0,44	0,016	0,022	0,014	0,018	0,61
shimmer local	0,055	0,030	0,054	0,021	0,71	0,040	0,027	0,036	0,022	0,37
shimmer apq3	0,029	0,018	0,029	0,012	0,76	0,021	0,014	0,019	0,012	0,45
shimmer apq5	0,033	0,020	0,032	0,013	0,75	0,022	0,012	0,020	0,012	0,59
shimmer apq11	0,044	0,018	0,043	0,015	0,76	0,029	0,014	0,027	0,013	0,38
shimmer dda	0,088	0,055	0,086	0,037	0,76	0,064	0,043	0,057	0,037	0,45
HNR	19,409	4,596	19,549	3,864	0,72	23,306	4,718	22,942	4,611	0,86
NHR	0,045	0,069	0,034	0,035	0,33	0,040	0,081	0,035	0,078	0,80

FIGURE 7.7 – Moyennes et écarts types des paramètres de phonation extraits lors des tâches de voyelles soutenues /a/ enregistrées avec le microphone professionnel. Les valeur-p sont issues du Welch t-test.

NHR) ont tendances à être plus élevés chez les hommes MP que chez les hommes sains, traduisant une voix plus éraillée. Pour les femmes, deux sujets sains ayant la voix particulièrement cassées ont un gros impact sur la moyenne de ces paramètres, ne permettant pas d'interprétation de ces moyennes. Concernant les paramètres plus liés aux tremblements, tels que les jitters (variations courts termes de Fo) et les shimmers (variations court terme de Int), on observe pour les deux sexes des moyennes plus élevées dans les groupes MP.

Bien que certaines tendances émergent de ces paramètres en faveur d'un timbre plus éraillé et plus tremblant chez les MP, les écarts ne sont pas significatifs. Ceci peut s'expliquer par le fait que ces paramètres sont moins altérés quand les MP sont sous traitements médicamenteux [Rusz et al., 2013b].

Cas des iRBD Concernant les iRBD, nous avons extraits les paramètres NHR pour évaluer l'éraillement du timbre, le jitter local pour évaluer le tremblement de la hauteur de la voix, et le shimmer local pour évaluer le tremblement de l'intensité . Les résultats obtenus avec le rappel de ces valeurs pour les hommes MP et sains sont présentés Figure 7.8.

Paramètres	MP		iRBD		sain		p (iRBD vs MP)	p (iRBD vs sain)
	moy	SD	moy	SD	moy	SD		
jitter local	0,010	0,012	0,012	0,019	0,008	0,005	0,17	0,09
shimmer local	0,055	0,030	0,064	0,036	0,054	0,021	0,47	0,21
nhr	0,045	0,069	0,057	0,092	0,034	0,035	0,42	0,12

FIGURE 7.8 – Moyennes et écarts types des paramètres de phonation lors des tâches de voyelles soutenues /a/ enregistrées avec le microphone professionnel. Comparaison entre les groupes hommes MP, iRBD et sains. Les valeur-p sont issues du Welch t-test.

Nous pouvons constater que ces 3 paramètres sont en moyenne plus altérés chez les iRBD que chez les MP, bien que les écarts ne soient toujours pas significatifs, ceci pouvant être lié au fait que les iRBD sont non traités pharmacologiquement pour la maladie de Parkinson.

7.2.3 Pauses

7.2.3.1 Répartition des pauses lors du monologue

Nous avons dans un premier temps étudié les pauses effectuées lors du monologue enregistré avec le microphone professionnel. Les résultats obtenus sont détaillés Figure 7.9. Nous pouvons constater que les MP et les sujets sains effectuent en moyenne le même nombre de pauses par seconde parlée, mais ils n'effectuent pas des pauses de même longueur. Que ce soit chez les hommes ou chez les femmes, les MP effectuent moins de pauses courtes (entre 200 et 500ms). De plus les hommes MP font plus de pauses longues (supérieures à 1s) que les hommes sains. Ces résultats sont confirmés par la durée médiane des pauses qui est plus grande chez les MP que chez les sains (de manière significative chez les hommes). La diminution du nombre de pauses courtes et l'augmentation du nombre de pauses longues (en tout cas pour les hommes) sont bien cohérentes avec la perception de débit saccadé que l'on peut avoir de la voix parkinsonienne.

	Hommes					Femmes				
	MP		sain		valeur-p	MP		sain		valeur-p
	moy	SD	moy	SD		moy	SD	moy	SD	
Nb pauses 200 à 500ms	0,228	0,216	0,296	0,171	0,068	0,262	0,172	0,365	0,266	**0,039**
Nb pauses 500 à 1s	0,216	0,108	0,224	0,111	0,69	0,200	0,088	0,216	0,155	0,55
Nb pauses > 1s	0,185	0,200	0,093	0,097	**3,8E-03**	0,072	0,094	0,082	0,282	0,83
Nb pauses > 200ms	0,629	0,398	0,614	0,261	0,82	0,537	0,245	0,668	0,683	0,25
Médiane pauses > 200ms	0,735	0,272	0,538	0,199	**3,3E-05**	0,548	0,192	0,456	0,128	**0,011**

FIGURE 7.9 – Moyennes et écarts types du nombre de pauses par seconde parlée lors du monologue enregistré avec le microphone professionnel. Les pauses sont comptabilisées en plusieurs tranches selon leurs durées, la durée médiane est aussi calculée. Les valeurs-p sont issues du Welch t-test.

7.2.3.2 Influence de la tâche

Nous avons voulu tester l'influence du type de tâche en effectuant la même analyse sur la lecture et la comparer aux résultats obtenus avec le monologue. Les résultats obtenus sont présentés Figure 7.10. Nous pouvons constater, comme pour le monologue, une diminution des pauses courtes chez les MP, pour les hommes et pour les femmes. Cependant, nous ne retrouvons pas l'augmentation des pauses longues chez les hommes MP, et notons même une diminution des pauses longues chez les MP femmes. D'une manière générale l'utilisation des pauses longues (portant beaucoup d'information syntaxique et stylistique) chez les sujets sains diffèrent suivant le type d'exercice vocal : on n'utilise pas autant de pauses lors d'un discours préparé, d'une conversation, d'une lecture ou du récit improvisé de sa journée. D'après nos résultats, l'augmentation des pauses longues chez les hommes MP semble être restreinte à certains exercices vocaux particuliers, comme le récit non préparé de sa journée, tâche qui implique une composante cognitive de mémorisation. A l'inverse, la diminution des pauses courtes chez les MP (et par la même occasion l'augmentation de la durée médiane) semble être persistant quel que soit le type d'exercice vocal.

7.2.3.3 Influence du microphone

Nous avons analysé le nombre moyen de pauses courtes et de pauses longues ainsi que la durée médiane extraits à partir cette fois des enregistrements issus du microphone interne de l'ordinateur et à partir des enregistrements téléphoniques pour évaluer l'effet du type de microphone sur ces paramètres. Les résultats sont présentés Figure 7.11. Nous retrouvons avec le microphone de l'ordinateur et le téléphone une diminution des pauses courtes et une augmentation de la médiane chez les MP (hommes et femmes). Une augmentation du nombre de pauses

	Hommes					Femmes				
Colonne1	MP		sain		valeur-p	MP		sain		valeur-p
	moy	SD	moy	SD		moy	SD	moy	SD	
Nb pauses 200 à 500ms	0,249	0,134	0,311	0,171	**0,027**	0,253	0,124	0,314	0,113	**0,021**
Nb pauses > 1s	0,089	0,086	0,090	0,087	0,94	0,033	0,055	0,068	0,073	**0,017**
Médiane pauses > 200ms	0,620	0,155	0,584	0,135	0,20	0,565	0,119	0,522	0,108	0,085

FIGURE 7.10 – Moyennes et écarts types du nombre de pauses par seconde parlée lors de la lecture (texte et dialogue) enregistrée avec le microphone professionnel. Les pauses sont comptabilisées en plusieurs tranches selon leurs durées, la durée médiane est aussi calculée. Les valeurs-p sont issues du Welch t-test.

longue est également retrouvée pour les hommes MP. En regardant les résultats pour chaque sujet, on s'aperçoit que certains bruits ont des répercussions sur l'intensité maximale, ce qui entraine une surdétection des silences (le seuil d'intensité des silences étant calculé par rapport à l'intensité maximale) et crée des outliers dans le nombre de pauses. De par leur localisation, le microphone interne de l'ordinateur est plus sensible aux bruits de type coup dans la table, et le microphone professionnel aux bruits de type toux. Ceci explique les légères différences qu'on peut observer dans les nombres de pauses détectées. Le microphone professionnel semble légèrement plus discriminant que le microphone interne de l'ordinateur avec ce type de paramètres. Quant au téléphone, les écarts entre les groupes sont moins significatifs, ce qui pourrait être en partie dû au nombre un peu plus faible de sujets dans la base de données téléphoniques.

		Hommes					Femmes				
Micro	Paramètres	MPh		sainh		valeur-p	MPf		sainf		valeur-p
		moy	SD	moy	SD		moy	SD	moy	SD	
micro pro	Nb pauses 200 à 500ms	0,228	0,216	0,296	0,171	0,068	0,262	0,172	0,365	0,266	**0,039**
	Nb pauses > 1s	0,185	0,200	0,093	0,097	**3,8E-03**	0,072	0,094	0,082	0,282	0,83
	Médiane pauses > 200ms	0,735	0,272	0,538	0,199	**3,3E-05**	0,548	0,192	0,456	0,128	**0,011**
micro ordi	Nb pauses 200 à 500ms	0,305	0,251	0,454	0,346	**6,8E-03**	0,398	0,213	0,497	0,478	0,23
	Nb pauses > 1s	0,247	0,445	0,125	0,138	0,068	0,109	0,132	0,157	0,498	0,55
	Médiane pauses > 200ms	0,694	0,271	0,510	0,171	**4,9E-05**	0,493	0,166	0,444	0,152	0,17
téléphone	Nb pauses 200 à 500ms	0,375	0,217	0,467	0,247	0,056	0,446	0,214	0,482	0,213	0,51
	Nb pauses > 1s	0,211	0,191	0,161	0,156	0,190	0,105	0,128	0,108	0,139	0,92
	Médiane pauses > 200ms	0,649	0,188	0,539	0,190	**6,30E-03**	0,494	0,189	0,470	0,112	0,57

FIGURE 7.11 – Moyennes et écarts types du nombre de pauses par seconde parlée, et durée médiane des pauses, lors du monologue. Comparaisons entre les enregistrements issus du microphone professionnel, ceux issus du microphone interne de l'ordinateur et les enregistrements téléphoniques. Les valeurs-p sont issues du Welch t-test.

7.2.3.4 Cas des iRBD

Nous avons analysé, cette fois pour le groupe iRBD, les paramètres les plus discriminants d'après les analyses précédentes sur les MP et les sains, soit le nombre de pauses courtes lors du monologue et de la lecture, et le nombre de pauses longues ainsi que la durée médiane lors du monologue. La Figure 7.12 regroupe les résultats des hommes iRBD, MP et sains, ainsi que les valeurs-p issues du t-test, correspondant aux enregistrements issus du microphone professionnel. La Figure 7.13 illustre plus particulièrement la répartition des durées médianes lors du monologue.

Nous constatons, quel que soit le paramètre, que la valeur moyenne pour le groupe iRBD se situe toujours entre celle du groupe MP et celle du groupe sain. Ce résultat montre que la répartition des pauses chez les iRBD connait les mêmes changements que chez les MP mais en moins prononcés, ce qui semble indiquer que le phénomène de débit saccadé, tout comme la monotonie de la voix, apparaît dès le stade prodromique de la maladie de Parkinson.

Tâches	Paramètres	MP		iRBD		sain		p (iRBD vs MP)	p (iRBD vs sain)
		moy	SD	moy	SD	moy	SD		
Monologue	Nb pauses 200 à 500ms	0,228	0,216	0,259	0,146	0,296	0,171	0,41	0,28
	Nb pauses > 1s	0,185	0,200	0,163	0,184	0,093	0,097	0,58	**0,024**
	Médiane pauses > 200ms	0,735	0,272	0,650	0,284	0,538	0,199	0,12	**0,032**
Lecture	Nb pauses 200 à 500ms	0,249	0,134	0,284	0,135	0,311	0,171	0,19	0,41

FIGURE 7.12 – Moyennes et écarts types du nombre de pauses par seconde parlée lors du monologue et de la lecture, enregistrés avec le microphone professionnel. Les valeur-p sont issues du Welch t-test comparant les groupes hommes iRBD et MP d'une part et les groupes hommes iRBD et sain d'autre part.

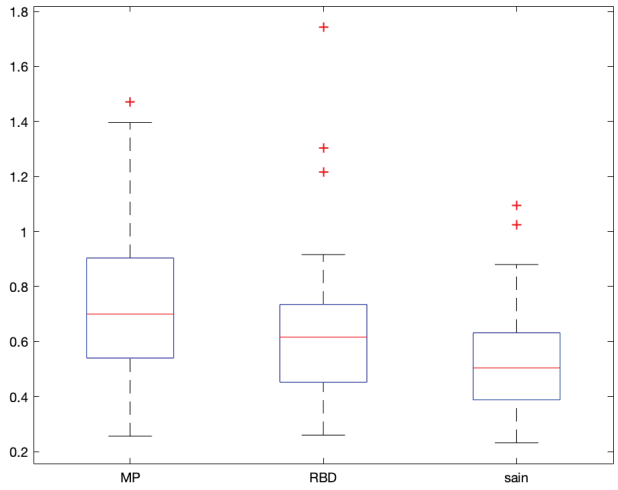


FIGURE 7.13 – Boites à moustaches représentant la répartition des durées médianes des pauses des hommes MP, iRBD et sains, extraites lors du monologue, enregistré avec le microphone professionnel. Les traits rouges représentent les médianes, les rectangles bleus les percentiles 25 et 75, et les croix rouges les outliers.

7.2.4 Rythme

7.2.4.1 Analyses du rythme lors des différentes tâches

Nous avons extrait les paramètres RSD et Ecart définis partie 7.1.4 à partir des tâches de répétitions lentes des syllabes /pa/, /kou/ et /pa kou/ (consistant en une répétition alternée des syllabes /pa/ et /kou/), enregistrées avec le microphone professionnel. Les résultats obtenus sont détaillés Figure 7.14.

Nous pouvons constater que, quelles que soient les syllabes à répéter (/pa/, /kou/ ou /pa kou/), le groupe MP a un écart type relatif (RSD) des durées d'intervalles plus élevé que le groupe sain, ceci étant vrai aussi bien pour les hommes que pour les femmes. Ceci signifie qu'en moyenne le rythme effectué par un sujet MP connaît plus de variabilité que celui d'un sujet sain. De plus nous pouvons noter que la variabilité est encore plus grande pour la tâche de syllabes alternées, ce qui va dans le sens de ce qu'avaient constaté les auteurs de [Skodda et al., 2013] à propos de patients parkinsoniens d'un stade un peu plus avancé. La moyenne des deux tâches /pa kou/ permet de lisser les quelques anomalies de détection des syllabes qui peuvent être rencontrées, augmentant la différence de RSD entre hommes MP et sains. La moyenne des 4 tâches semble légèrement moins discriminante que la moyenne des tâches /pa kou/.

		Hommes					Femmes				
Tâches	Paramètres	MP		sain		valeur-p	MP		sain		valeur-p
		moy	SD	moy	SD		moy	SD	moy	SD	
/pa/	RSD	0,087	0,078	0,062	0,045	0,053	0,099	0,092	0,091	0,091	0,69
	Ecart	0,149	0,129	0,114	0,102	0,12	0,157	0,111	0,135	0,115	0,38
/kou/	RSD	0,085	0,074	0,058	0,036	**0,020**	0,088	0,064	0,078	0,056	0,45
	Ecart	0,124	0,096	0,098	0,092	0,13	0,126	0,119	0,109	0,084	0,45
/pa kou/	RSD	0,112	0,100	0,073	0,055	**0,015**	0,125	0,112	0,078	0,049	**0,013**
	Ecart	0,231	0,167	0,193	0,133	0,19	0,252	0,168	0,172	0,113	**0,012**
moy /pa kou/	RSD	0,118	0,090	0,076	0,053	**4,3E-03**	0,130	0,108	0,091	0,052	**0,036**
	Ecart	0,225	0,166	0,183	0,135	0,15	0,251	0,151	0,178	0,120	**0,016**
moy tout	RSD	0,102	0,069	0,068	0,036	**2,3E-03**	0,112	0,084	0,088	0,050	0,11
	Ecart	0,181	0,121	0,145	0,095	0,083	0,196	0,112	0,150	0,088	**0,038**

FIGURE 7.14 – Moyennes et écarts types des paramètres RSD (écart type relatif des durées séparant 2 syllabes consécutives) et Ecart (différence entre la moyenne de ces durées et la durée des intervalles dans l'exemple). Ces paramètres ont été extraits à partir des enregistrements du microphone professionnel, lors des tâches de répétition lente des syllabes /pa/, /kou/ et une tâche /pa kou/ d'alternance entre les syllabes /pa/ et /kou/. Les moyennes des paramètres extraits lors des 2 deux tâches /pa kou/ sont calculées ainsi que la moyenne sur les 4 tâches de répétitions lentes. Les valeurs-p sont issues du Welch t-test.

Concernant le paramètre Ecart, nous pouvons constater que pour toutes les tâches, aussi bien chez les hommes que chez les femmes, il est plus élevé chez les MP que chez les sains. Ceci indique que les sujets MP ont plus de mal à reproduire globalement un rythme imposé (sans tenir compte des variabilités de rythme au cours de la tâche). Nous retrouvons également que la différence entre les MP et les sains concernant ce paramètre est plus élevée pour la tâche de répétition alternée /pa kou/ que pour les tâches de répétitions non alternées.

7.2.4.2 Influence des microphones

Nous avons ensuite extrait les paramètres rythmiques à partir des enregistrements issus du microphone interne de l'ordinateur et à partir des enregistrements téléphoniques. Les résultats concernant les tâches de répétitions lentes /pa kou/ sont présentés Figure 7.15. Concernant les enregistrements du microphone professionnel et du microphone interne de l'ordinateur, les paramètres sont moyennés sur les 2 tâches /pa kou/. Concernant les enregistrements téléphoniques, ils sont extraits de la tâche /pa kou/ de chaque session, puis moyennés sur l'ensemble des sessions.

		Hommes					Femmes				
Micro	Paramètres	MP		sain		valeur-p	MP		sain		valeur-p
		moy	SD	moy	SD		moy	SD	moy	SD	
micro pro	RSD	0,118	0,090	0,076	0,053	**4,3E-03**	0,130	0,108	0,091	0,178	**0,036**
	Ecart	0,225	0,166	0,183	0,135	0,15	0,251	0,151	0,052	0,120	**0,016**
micro ordi	RSD	0,135	0,108	0,092	0,076	**0,017**	0,139	0,109	0,103	0,065	0,068
	Ecart	0,223	0,162	0,183	0,129	0,15	0,248	0,148	0,178	0,117	**0,019**
téléphone	RSD	0,117	0,060	0,098	0,071	0,16	0,103	0,058	0,093	0,044	0,47
	Ecart	0,290	0,159	0,262	0,176	0,42	0,333	0,187	0,331	0,210	0,96

FIGURE 7.15 – Moyennes et écarts types des paramètres RSD et Ecart extraits et moyennés à partir des tâches /pa kou/. Comparaison entre les enregistrements issus du microphone professionnel, du microphone interne de l'ordinateur et des téléphones. Les valeurs-p sont issues du Welch t-test.

Nous pouvons constater une augmentation de la variabilité du rythme (RSD) et une augmentation de l'écart par rapport au rythme imposé (Ecart) pour les groupes MP (hommes et femmes), avec les trois types d'enregistrements. Cependant nous pouvons noter des écarts un peu moins marqués pour les enregistrements issus du microphone de l'ordinateur, comparés au

microphone professionnel. Ceci est dû à une détection des syllabes légèrement meilleure pour les enregistrements issus du microphone professionnel. Une optimisation différente de la VAD pourrait être faite pour correspondre plus spécifiquement aux caractéristiques des enregistrements issus du microphone de l'ordinateur. De même les écarts entre les groupes, bien qu'existants, sont moins marqués pour les enregistrements téléphoniques. Ceci semblerait être dû en partie à une VAD moins optimisée et également un moins bon respect des consignes.

7.2.4.3 Cas des iRBD

Enfin nous avons extrait les paramètres rythmiques chez les sujets hommes iRBD, et comparé les valeurs obtenues avec celles des groupes hommes MP et sains, cf. Figure 7.16.

Paramètres	MP		iRBD		sain		p (iRBD vs MP)	p (iRBD vs sain)
	moy	SD	moy	SD	moy	SD		
RSD	0,118	0,090	0,117	0,089	0,076	0,053	0,98	**8,2E-03**
Ecart	0,225	0,166	0,171	0,148	0,183	0,135	0,085	0,68

FIGURE 7.16 – Moyennes et écarts types des paramètres RSD et Ecart extraits et moyennés à partir des tâches /pa kou/ enregistrées avec le microphone professionnel. Les valeur-p sont issues du Welch t-test comparant les groupes hommes iRBD et MP d'une part et les groupes hommes iRBD et sain d'autre part.

Nous pouvons constater de façon assez inattendue que le groupe des iRBD a une variabilité relative du rythme (*RSD*) similaire au groupe MP en moyenne, et un écart moyen absolu par rapport au rythme de l'exemple (*Ecart*) similaire au groupe sain. D'après ce résultat, nous pouvons émettre l'hypothèse que la capacité à garder un rythme constant, et la capacité à répéter un rythme imposé, feraient intervenir des processus physiologiques différents, dont le premier serait altéré plus tôt que le deuxième, dans la maladie de Parkinson.

7.3 Classification avec SVM

A partir des analyses de variance précédentes, nous avons choisi un ensemble réduit de paramètres globaux pour effectuer une classification SVM. Pour les enregistrements effectués avec le microphone professionnel et le microphone de l'ordinateur, nous avons choisi un ensemble de 6 paramètres discriminants, en équilibrant les domaines de la voix liés à la prosodie, aux pauses et à la capacité de suivre un rythme. Nous avons choisi les paramètres suivants :

- *SD log Fo*, moyenné sur les tâches de lecture, répétition de phrase et monologue
- *SD log Fo*, extrait seulement à partir du dialogue
- le nombre de pauses $\in$ [200ms, 500ms], lors du monologue
- la médiane des pauses $>$ 200ms, lors du monologue
- *RSD*, moyenné sur les deux tâches de répétions lentes des syllabes /pa kou/
- *Ecart*, moyenné sur les deux tâches de répétions lentes des syllabes /pa kou/

Concernant les enregistrements téléphoniques, un paramètre s'est avéré nettement plus discriminant que les autres : *SD log Fo*, extrait à partir des répétitions de phrases et du monologue. Nous avons gardé seulement ce paramètre pour la classification.

Nous avons ensuite entrainé un SVM à marges souples (cf. partie 3.1.2.2) pour la classification MP vs sain à partir de ces paramètres. Tout comme pour les analyses précédentes, nous avons séparé les hommes des femmes pour l'entraînement du modèle et la classification. Nous avons utilisé le logiciel matlab pour les entraînements des SVM et la classification des sujets tests.

Nous avons pris une fonction noyau linéaire, considéré des *a priori* de classes égaux, et normalisé les paramètres afin de leur donner un poids équivalent quel que soit leur ordre de grandeur.

Nous avons commencé par entraîner un SVM sur toute la base de données (hommes et femmes séparément), en faisant varier les valeurs des hyperparamètres liés à la pénalité de violation de marges, et liés à l'échelle de la fonction noyau. Les valeurs optimales trouvées ont été gardées pour les entraînements et tests suivants.

Tout comme pour les analyses à partir des MFCC présentées dans les chapitres précédents, nous avons utilisé une méthode ensembliste pour la classification.

Nous avons opté pour l'agrégation de 10 validations croisées de type 10-fold. Nous avons préféré ce partitionnement au *repeated random subsampling* (utilisé pour les analyses MFCC-GMM), n'ayant pas ici de nécessité particulière à avoir le même nombre de sujets d'entraînement MP que sain.

Pour chaque 10-fold, nous avons partitionné nos sujets aléatoirement en 10 "plis" disjoints et effectué 10 runs de classification. Pour chaque run, nous avons entrainé un SVM à partir des sujets de 9 plis puis testé les sujets du pli restant. A l'issu des 10 runs d'un 10-fold, chaque sujet a été testé exactement une fois. A la fin des 10 10-fold, chaque sujet a ainsi été testé 10 fois. L'agrégation a consisté à moyenner les 10 scores de classification des sujets. Les EER et les courbes DET ont ensuite été calculées à partir de ces scores moyens.

7.4 Résultats classification MP vs sain

7.4.1 Résultats avec les trois types de microphones

Les résultats des classifications MP vs sain à partir des SVM sont présentés dans le Tableau 7.1. Les trois types de microphones (professionnel, ordinateur et téléphone) ont été analysés. Les courbes DET de la Figure 7.17 illustrent plus spécifiquement les performances de la classification MP vs sain, chez les hommes et chez les femmes, à partir du microphone professionnel.

microphone	paramètres	EER hommes	EER femmes
micro pro	6 param	22 ± 6%	32 ± 7%
micro ordi	6 param	22 ± 6%	30 ± 7%
téléphone	SD log Fo	27 ± 6%	33 ± 7%

TABLE 7.1 – Comparaison des EER (moyenne ± écart type) obtenus pour la classification agrégée des SVM, MP vs sain, à partir des paramètres globaux, avec les 3 types d'enregistrements (issus du microphone professionnel, du microphone interne de l'ordinateur et des téléphones des sujets). L'ensemble des 6 paramètres comprend *SD log Fo* (extrait des tâches de lecture, répétitions de phrases et monologue), *SD log Fo* extrait de la lecture du dialogue, le nombre de pauses $\in$ [200ms, 500ms] et la médiane des pauses $>$ 200ms (extraits du monologue) et les variations du rythme *RSD* et *Ecart* (extraits des répétitions lentes des syllabes /pa kou/). Pour les enregistrements téléphoniques, seul *SD log Fo* extrait à partir des répétitions de phrases et du monologue a été gardé pour la classification.

Concernant les enregistrements issus du microphone professionnel et du microphone interne de l'ordinateur, nous n'observons pas de différences dans les performances de ce classifieur. Nous avons obtenu un EER de 22% pour la classification des hommes et un EER d'environ 31% pour la classification des femmes. Nous constatons une meilleure détection de MP chez les hommes que chez les femmes, confirmant les différences de performances déjà constatées à partir des analyses faites dans les deux chapitres précédents. Ces classifications SVM ont été effectuées

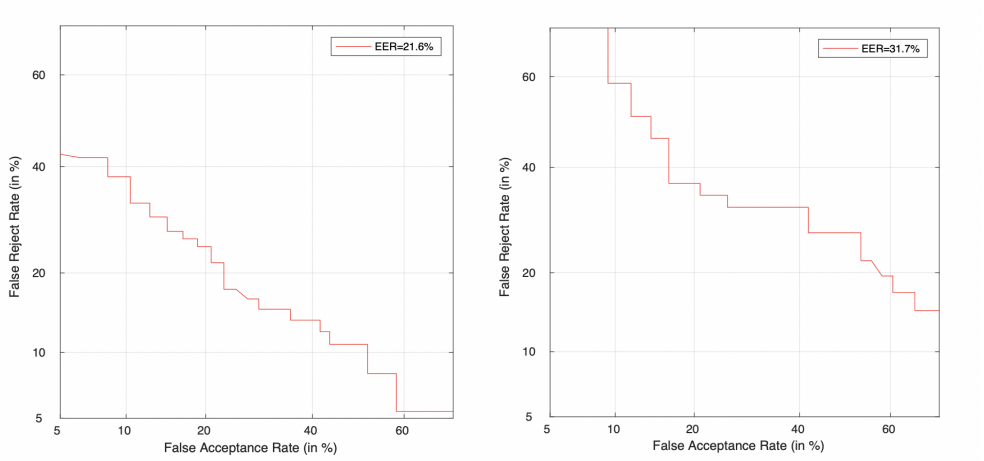


FIGURE 7.17 – Courbes DET résultant de la classification agrégée des SVM, MP vs sain, à gauche pour les hommes, à droite pour les femmes. Les classifications SVM ont été effectuées à partir des paramètres globaux, extraits des enregistrements du microphone professionnel.

avec l'ensemble réduit des 6 paramètres globaux présenté dans la section précédente, extraits à partir d'environ 3 min de parole par sujet.

Pour les enregistrements téléphoniques, plusieurs sets de paramètres globaux ont été testés, mais les meilleurs résultats ont été obtenus avec le paramètre SD log Fo extrait à partir des répétitions de phrases et du monologue. Ceci est cohérent avec les analyses de variances précédentes montrant que ce paramètre était de loin le plus discriminant avec ce type d'enregistrements. Nous avons obtenu un EER de 27% pour la détection de MP chez les hommes et 33% pour les femmes. Toutes les sessions téléphoniques ont été utilisées pour calculer le paramètre SD log Fo, ce qui fait une moyenne de 6 min de parole par sujet pour l'extraction de ce paramètre.

Enfin, afin de vérifier que les différents lieux d'enregistrements n'ont pas eu d'effet significatif sur les classifications, nous avons comparé les performances des sujets enregistrés à l'hôpital à celles des sujets enregistrés hors hôpital. Les différences avec les performances calculées sur tous les sujets n'ont pas dépassé 2%, et les t-tests n'ont pas montré d'écart significatif entre les différents lieux d'enregistrements.

7.4.2 Comparaison modèle agrégé avec modèle simple

Tout comme les analyses faites dans les deux chapitres précédents, nous avons également évalué les performances du modèle simple (non agrégé). Pour estimer les performances généralisées, nous avons moyenné les EER obtenus pour chacun des 10-fold. Les résultats obtenus pour les différents types d'enregistrements sont présentés Tableau 7.2.

Nous n'observons ici pas de différences significatives avec les performances du modèle agrégé. Nous pouvons conclure que l'agrégation des modèles n'est pas spécialement nécessaire pour ce type de classifieur.

microphone	paramètres	EER hommes	EER femmes
micro pro	6 param	22 ± 6%	33 ± 7%
micro ordi	6 param	21 ± 6%	33 ± 7%
téléphone	SD log Fo	27 ± 6%	33 ± 7%

TABLE 7.2 – Comparaison des EER (moyenne ± écart type) obtenus pour la classification SVM (non agrégée), MP vs sain, à partir des paramètres globaux, avec les 3 types d'enregistrements. L'ensemble des 6 paramètres comprend *SD log Fo* (extrait des tâches de lecture, répétitions de phrases et monologue), *SD log Fo* extrait de la lecture du dialogue, le nombre de pauses ∈ [200ms, 500ms] et la médiane des pauses > 200ms (extraits du monologue) et les variations du rythme *RSD* et *Ecart* (extraits des répétitions lentes des syllabes /pa kou/). Pour les enregistrements téléphoniques, seul *SD log Fo* extrait à partir des répétitions de phrases et du monologue a été gardé pour la classification.

7.5 Résultats classification iRBD vs sain

Nous avons effectué les mêmes types de classification avec cette fois les iRBD. Les résultats obtenus à partir du microphone professionnel sont présentés Tableau 7.3.

paramètres	iRBD vs sain	iRBD $moteur^+$ vs sain
6 param	38 ± 8%	33 ± 7%

TABLE 7.3 – EER (moyenne ± écart type) obtenus pour la classification agrégée des SVM, iRBD vs sain, à partir de l'ensemble réduit des 6 paramètres globaux, enregistrés avec le microphone professionnel. A savoir *SD log Fo* (extrait des tâches de lecture, répétitions de phrases et monologue), *SD log Fo* extrait de la lecture du dialogue, le nombre de pauses ∈ [200ms, 500ms] et la médiane des pauses > 200ms (extraits du monologue) et les variations du rythme *RSD* et *Ecart* (extraits des répétitions lentes des syllabes /pa kou/). Comparaison avec la classification des iRBD $moteur^+$ vs sain.

Nous constatons un EER de 38% pour la classification des iRBD vs sain. Cette performance est améliorée si nous nous limitons aux iRBD $moteur^+$ (EER de 33%). Nous considérons comme iRBD $moteur^+$ les iRBD dont le score UPDRS III est supérieur ou égal à 14 (cf. partie 4), c'est à dire des sujets pas encore diagnostiqués comme MP mais qui commencent à développer des symptômes moteurs.

7.6 Conclusion sur les analyses avec les paramètres globaux

Pour résumer, nous avons, dans ce chapitre, analysé des paramètres dits globaux, décrivant les différents changements observables dans la voix des sujets MP. Nous avons extrait des paramètres attraits à la prosodie, à la phonation, à l'utilisation des pauses, et à la capacité à suivre un rythme. Ces paramètres sont dits globaux car ils sont calculés à l'échelle de la tâche vocale et non à l'échelle des trames court termes (cas des MFCC).

Nous avons effectué des analyses de variance afin de savoir quels paramètres différaient de manière significative entre les groupes MP et sains, et quelles tâches vocales mettaient le mieux en valeur ces différences. Les paramètres qui se sont révélés les plus discriminants sont :

- *SD log Fo*, dont la diminution chez les MP traduit la monotonie de l'intonation. La diminution de la prosodie concerne les répétitions de phrases, le monologue, et particulièrement la lecture du dialogue à contenance émotionnelle.

- Le nombre de pauses $\in$ [200ms, 500ms] (réduit chez les MP) et la médiane des pauses supérieures à 200ms (allongée chez les MP) extraits du monologue. Les altérations de ces paramètres traduisent un débit de parole saccadé chez les MP.
- La variation relative du rythme (*RSD*) lors de la répétition lente de syllabes et l'*Ecart* moyen avec le rythme imposé. Ces paramètres sont augmentés chez les MP, traduisant une difficulté à garder un rythme constant et suivre un rythme imposé. Ces variations rythmiques sont majorées lors de la répétition alternée des syllabes /pa/ et /kou/, par rapport aux répétitions non alternées.

Pour tous ces paramètres on constate des différences entre les groupes MP et sain, que ce soit chez les hommes ou chez les femmes. Néanmoins les différences sont plus marquées chez les hommes, ce qui est cohérent avec les résultats des chapitres précédents.

Nous avons analysé ces paramètres à partir des enregistrements du microphone professionnel, du microphone interne de l'ordinateur et des enregistrements téléphoniques. Nous avons étudié, pour chaque paramètre, l'influence de ces types d'enregistrements. Nous avons globalement constaté peu de différences entre les résultats des enregistrements du microphone professionnel et ceux du microphone de l'ordinateur. Les paramètres extraits à partir des enregistrements téléphoniques se sont, quant à eux, montrés un peu moins discriminants, mais certaines différences restaient significatives.

Les variations de l'intensité et les paramètres attraits à la phonation, se sont révélés peu discriminants. Ceci peut s'expliquer par le fait que les altérations de ces deux types de paramètres sont connues pour être atténuées par les traitements dopaminergiques [Rusz et al., 2013b], or quasiment tous nos patients sont traités et ont été enregistrés en ON.

A partir de ces analyses nous avons choisi un ensemble réduit de 6 paramètres pour effectuer une classification, MP vs sain, de type SVM, avec une fonction noyau linéaire. L'ensemble de paramètres comprend : SD log Fo extrait dans deux conditions différentes, les deux paramètres liés aux pauses cités précédemment, et les deux paramètres rythmiques extraits lors des tâches /pa kou/. Nous avons obtenu comme résultat, que ce soit avec le microphone professionnel ou avec le microphone de l'ordinateur, un EER de 22% pour les hommes, et d'environ 31% pour les femmes, le tout à partir de 3 min de parole par sujet.

Concernant les enregistrements téléphoniques, les meilleurs résultats ont été obtenus, en considérant seulement le paramètre prosodique *SD log Fo*. Les EER correspondants sont de 27% pour les hommes et 33% pour les femmes. Les différentes sessions téléphoniques ont été analysées pour extraire ce paramètre, aboutissant à une quantité de paroles utilisée d'environ 6 min par sujet. Cette diminution des performances par rapport aux microphones professionnel et de l'ordinateur serait en partie due à une VAD moins optimisée et un moins bon respect des consignes (notamment pour la tâche de répétitions lentes).

Nous avons comparé les performances d'une méthode ensembliste (agrégation d'un 10 fois 10-fold) avec celles du modèle simple, et n'avons pas constaté de différence significative. L'apport de l'agrégation n'apparaît donc pas nécessaire pour ce type de classifieur.

Enfin concernant les iRBD, nous avons pu constater que les moyennes des paramètres liés à la prosodie, et ceux liés aux pauses, étaient à chaque fois comprises entre les moyennes des MP et des sains. Ceci va dans le sens d'une altération progressive et dès le stade prodromique de ces domaines vocaux. Concernant les paramètres rythmiques, seule la variabilité relative du rythme (*RSD*) semble être altérée chez les iRBD. Ceci pourrait signifier que la capacité à garder un rythme constant, et la capacité à répéter un rythme imposé, feraient intervenir des

processus physiologiques différents, dont le premier serait altéré plus tôt que le deuxième, dans le développement de la maladie de Parkinson.

La classification des iRBD par rapport aux sujets sains, avec le classifieur SVM (à partir du même ensemble de 6 paramètres) a conduit à un EER de 38% pour les enregistrements du microphone professionnel. Ces résultats sont améliorés si on considère seulement les iRBD $moteur^+$ (EER de 33%).

Chapitre 8

Fusion des classifieurs et résultats finaux de classification

Dans les chapitres précédents, nous avons utilisé trois méthodes de classifications différentes pour détecter la maladie de Parkinson : la méthode MFCC-GMM, les x-vecteurs et les paramètres globaux classés avec un SVM. Ces méthodes sont différentes par plusieurs aspects : le type de tâches vocales utilisées, les paramètres extraits appartenant à différents domaines phonétiques, l'échelle de temps d'analyse et le type de classifieur. Il est donc probable qu'elles contiennent des informations complémentaires quant au caractère parkinsonien ou non d'une voix. Afin de prendre en compte cette complémentarité, nous avons, dans ce chapitre fusionné ces trois méthodes de classification. Nous présenterons les résultats issus de cette fusion et ferons un bilan sur les résultats finaux de classification et les méthodes les plus appropriées suivant le genre et le type de microphone utilisé.

8.1 Fusion des classifieurs

8.1.1 Méthode de fusion

La méthode de fusion utilisée est une méthode standard à vote majoritaire. Pour chacune des trois méthodes de classification, nous avons gardé les caractéristiques donnant les meilleures performances, suivant le genre et le microphone utilisé. Les classifications des trois classifieurs ont été faites aux seuils EER, et nous avons considéré comme décision finale la décision majoritaire.

8.1.2 Cas des hommes enregistrés avec le microphone professionnel

Nous avons commencé par utiliser les données des hommes enregistrés avec le microphone professionnel. Pour la méthode MFCC-GMM nous avons considéré la combinaison des tâches donnant les meilleurs résultats, à savoir la lecture et les répétitions de phrases (avec des GMM spécifiques) et la tâche de diadococinésie /pataka/ (avec des GMM généraux). Pour la méthode x-vecteur, nous avons considéré les tâches de lecture et répétitions de phrases, et pour la classification nous avons gardé l'analyse discriminante linéaire (LDA) suivie d'une distance cosinus. Enfin pour la méthode avec les paramètres globaux, nous avons utilisé en entrée du SVM, l'ensemble restreint de paramètres décrit section 7.3, extraits à partir des tâches de lecture, répétitions de phrases, monologue et répétition lente de syllabes. Les taux de bonnes classifications (Acc) aux seuils EER sont rappelés Figure 8.1, et y est présentée la performance de fusion. La matrice de confusion est détaillée Figure 8.2.

On constate un taux de bonnes classification (Acc) de **89%** (Se=87%, Sp=92%) à l'issue de la fusion, soit une **amélioration de 6%** des performances par rapport au meilleur classifieur.

Paramètres et classifieurs	MFCC GMM	X-vecteur LDA + distance cos	Paramètres globaux SVM	Fusion
Tâches vocales	Lecture Répétition phrases DDK	Lecture Répétition phrases	Lecture Répétition phrases Monologue Répét lente de syllabes	Lecture Répétition phrases DDK Monologue Répét lente de syllabes
Durée	1min30	1min10	3min10	3min30
Champ phonétique	Articulation	Articulation	Prosodie Fluence Rythme	Articulation Prosodie Fluence Rythme
Acc	83%	78%	78%	**89%**

FIGURE 8.1 – Performances et caractéristiques de la classification des hommes MP vs sain, avec les méthodes MFCC-GMM (cf. chapitre 5), x-vecteur (cf. chapitre 6) et paramètres globaux (cf. chapitre 7), et fusion de ces 3 classifieurs par vote majoritaire. Les enregistrements considérés sont ceux du microphone professionnel.

	MP Prédit	Sain Prédit
MP réel	87% ± 4%	13% ± 4%
Sain réel	8% ± 4%	92% ± 4%

FIGURE 8.2 – Matrice de confusion correspondant à la fusion des trois classifieurs (MFCC-GMM, x-vecteur, paramètres globaux), utilisés pour classer les hommes MP vs sain, enregistrés avec le microphone professionnel.

8.1.3 Cas des hommes enregistrés avec le téléphone

Nous nous sommes ensuite intéressés à la fusion de ces trois méthodes de classification, appliquée aux données téléphoniques. Pour la méthode MFCC-GMM et les x-vecteurs, nous avons gardé les tâches qui donnaient les meilleurs résultats au téléphone, à savoir les tâches de diadococinésie (DDK) (la lecture étant absente des enregistrements téléphoniques). En ce qui concerne les paramètres globaux, nous avons gardé le paramètre le plus discriminant, le paramètre prosodique SD log Fo, extrait à partir des répétitions de phrases et du monologue. Les résultats sont présentés dans la Figure 8.3.

Contrairement aux enregistrements issus du microphone professionnel, nous n'observons pas d'amélioration avec la fusion. Les performances de fusion (Acc=75%) sont égales à celle du meilleur classifieur (MFCC-GMM). L'absence d'amélioration suite à la fusion pourrait être due aux performances plus faibles des trois classifieurs. En effet, meilleurs sont les classifieurs qu'on fusionne, plus la probabilité que la fusion améliore les résultats est élevée.

8.1.4 Cas des femmes

En ce qui concerne les femmes enregistrées avec le microphone professionnel, nous avons gardé les tâches de lecture et de répétitions de phrases pour la méthode MFCC-GMM, la tâche

de monologue pour les x-vecteurs et pour le SVM le même ensemble de paramètres globaux que pour les hommes, donc extraits à partir des tâches de lecture, répétitions de phrases, monologue et répétition lente de syllabes. Les résultats sont détaillés dans la Figure 8.3.

	MFCC-GMM	X-vecteurs LDA + dist cos	Param globaux SVM	Fusion
Micro pro hommes	Acc = 83%	Acc = 78%	Acc = 78%	**Acc = 89%** Sp=92% Se=87%
Téléphone hommes	**Acc = 75%**	Acc = 71%	Acc = 73%	Acc = 75% Sp=78% Se=73%
Micro pro femmes	Acc = 60%	**Acc = 70%**	Acc = 68%	Acc = 69% Sp=67% Se=71%

FIGURE 8.3 – Performances de classification MP vs sain des trois classifieurs (MFCC-GMM, x-vecteur, paramètres globaux) et de leur fusion. Comparaison entre microphone professionnel et téléphone ainsi qu'entre hommes et femmes.

Nous ne constatons pas d'amélioration des performances suite à la fusion, ce qui peut être dû aux performances plus faibles des classifieurs, avec notamment le classifieur MFCC-GMM dont les mauvaises performances peuvent nuire au résultat de fusion.

8.2 Résultats finaux de classification

En résumé, la fusion a nettement amélioré les performances de classification pour les hommes enregistrés avec le microphone professionnel. En revanche, elle n'a pas apporté d'amélioration pour les enregistrements téléphoniques ni pour les femmes, ceci pouvant être dû aux performances plus faibles et plus hétérogènes des classifieurs. Ainsi, suivant le type de microphone utilisé et le genre, la méthodologie d'analyse optimale et les meilleures tâches vocales ne seront pas les mêmes. Un bilan des meilleurs résultats obtenus par type de microphone et par genre, avec pour chaque catégorie, les tâches utilisées et la méthode de classification, est présenté Figure 8.4.

	Micro professionnel				Micro ordinateur				Téléphone			
	durée	tâches	méthode	Acc	durée	tâches	méthode	Acc	durée	tâches	méthode	Acc
hommes	6min	tout sauf /a/	fusion	**89%**	4min	monol-lec-rythme	globale	**78%**	5min	DDK	MFCC-GMM	**75%**
femmes	1min	monol	x-vect	**70%**	4min	monol-lec-rythme	globale	**70%**	5min	monol	x-vect	**67%**

FIGURE 8.4 – Tableau récapitulatif des performances finales obtenues suivant le genre et le microphone utilisé. Détail des tâches et des méthodes de classification les plus appropriées par catégorie.

Nous pouvons constater que, quelque soit le type de microphone, les hommes ont de meilleures performances de classification que les femmes. De plus quelque soit le genre, le microphone professionnel a tendance à donner de meilleurs résultats que le microphone de l'ordinateur qui est lui-même plus performant que le téléphone.

Concernant les enregistrements du microphone professionnel, la fusion des trois méthodes de classification est à privilégier pour les hommes. Par contre au vu des plus faibles perfor-

mances obtenues avec le téléphone, il est préférable de prendre pour celui-ci le meilleur classifieur (MFCC-GMM) plutôt que la fusion. Pour les femmes, les faibles performances du classifieur MFCC-GMM (causées par la grande variabilité des MFCC des femmes) nuit à la fusion. Cette variabilité est réduite avec la méthode des x-vecteurs qui utilise une LDA, ce qui rend cette dernière méthode particulièrement adaptée à la classification des femmes. Enfin la méthode des paramètres globaux semble particulièrement adaptée au microphone de l'ordinateur pour les deux sexes. En effet nous avions vu que la fonction de débruitage actif et la distance accrue entre le microphone et la bouche ne semblaient pas altérer l'extraction de ce type de paramètres.

Chapitre 9

Corrélation des paramètres voix avec la neuroimagerie et la clinique

Les analyses sur les classifications MP débutant vs sain répondaient à un objectif d'aide au diagnostic précoce par l'analyse vocale. Pour répondre à l'objectif complémentaire d'aide au suivi de l'évolution de MP par l'analyse de la voix, nous avons analysé s'il y avait des corrélations entre des paramètres vocaux et l'état d'avancement de la maladie, quantifiée par des résultats de neuroimagerie, et le score moteur de l'UPDRS III. Pour cela nous avons essayé de prédire les résultats du DatScan, de l'Imagerie par Résonance Magnétique (IRM) sensible à la neuromélanine (NM) et de l'UPDRS III à partir d'un ensemble restreint de paramètres vocaux.

Au moment de cette analyse, tous les sujets analysés précédemment n'avaient pas encore été enregistrés, donc nous allons présenter les résultats obtenus à partir d'une partie réduite de notre base de données voix, composées de 142 sujets de la base ICEBERG, cf. [Jeancolas et al., 2019b]. La moyenne d'âge de ces sujets est de 62.8 ± 9.4 ans. Parmi ces sujets, 103 sont MP (68 hommes et 35 femmes) et 39 sont des sujets sains (17 hommes et 22 femmes). Tous les sujets MP étaient au niveau 2 sur l'échelle de Hoehn et Yahr [Hoehn and Yahr, 1967] avec une durée d'évolution de la maladie depuis le diagnostic de 32.6 ± 17.9 mois. Les sujets sains étaient tous à 0 sur l'échelle de Hoehn et Yahr.

Nous commencerons ce chapitre par un rapide état de l'art sur les corrélations des perturbations vocales avec la neuroimagerie et la clinique. Ensuite nous présenterons les données de neuroimagerie du DatScan et d'IRM sensible à la NM, ainsi que les données cliniques utilisées pour quantifier l'évolution de la maladie. Nous détaillerons ensuite les paramètres vocaux globaux que nous avons retenus pour les corrélations. Enfin nous expliquerons les modèles de régression que nous avons utilisés pour la prédiction des résultats de neuroimagerie et des scores moteurs à partir des paramètres vocaux, et donnerons les performances obtenues.

9.1 Etat de l'art sur les corrélations des perturbations vocales avec la neuroimagerie et la clinique

Plusieurs études se sont penchées sur l'utilisation de l'analyse vocale pour prédire le stade d'évolution de MP, caractérisé par le score clinique d'UPDRS III. Différents modèles de régression ont été utilisés, comme des forêts d'arbres décisionnels [Tsanas et al., 2011,Halawani and Ahmad, 2012], des SVM [Orozco-Arroyave et al., 2016b], des GPR (Gaussian Processes Regression) ainsi que des DNN [Grosz et al., 2015]. La prédiction du score moteur de l'UPDRS III à partir de paramètres vocaux, a d'ailleurs fait l'objet d'un challenge Interspeech en 2015 [Schuller et al., 2015].

Concernant les corrélats en neuroimagerie des perturbations vocales dans la maladie de Parkinson, une dizaine d'études se sont penchées sur le sujet, via l'analyse d'IRMf et de TEP prises pendant que les sujets effectuaient des tâches vocales. Les auteurs de [Sachin et al., 2008] et [Rektorova et al., 2012] ont mis en évidence une altération du circuit striato-cortical et des principales régions cérébrales motrices (cortex moteur, orofacial et cervelet). [Pinto et al., 2014] et [Narayana et al., 2009] ont montré l'existence d'une compensation avec une plus grande participation des cortex prémoteurs et préfrontaux (AMS, cortex moteur supérieur, DLPFC). [Sachin et al., 2008] a également mis en évidence une réorganisation fonctionnelle avec un recrutement supplémentaire des régions temporales.

Cependant la corrélation entre des paramètres vocaux, résultant d'une analyse acoustique, et les modifications que l'on peut observer en neuroimagerie a été très peu étudiée. Les auteurs de [Narayana et al., 2010] ont étudié le lien entre l'intensité de la voix et les données d'imagerie TEP sur 10 sujets, et les auteurs de [Rektorova et al., 2012] ont étudié le lien entre des paramètres vocaux (fréquence fondamentale et intensité) et des données d'IRM fonctionnelle sur 17 sujets.

À notre connaissance, la corrélation entre des paramètres vocaux et les changements dans le système dopaminergique, utilisant soit l'imagerie des transporteurs dopaminergiques (DAT) soit l'imagerie IRM sensible à la neuromélanine, n'avait quant à elle jamais été étudiée.

9.2 Données de neuroimagerie et paramètres cliniques

9.2.1 Analyse du DatScan

La 123-I Ioflupane tomoscintigraphie d'émission monophotonique (DatScan) est une imagerie des transporteurs dopaminergiques, qui est connue pour mettre en évidence l'atteinte du striatum dans la maladie de Parkinson en montrant une diminution de ces derniers dans cette région. Cette technologie, dérivée de la tomoscintigraphie d'émission monophotonique, utilise le traceur Ioflupane (I-123), qui se fixe spécifiquement sur les transporteurs présynaptiques de la dopamine du striatum (région intervenant dans le contrôle moteur). Dans la maladie de Parkinson, on constate avec cette imagerie une hypofixation du traceur au niveau du striatum, notamment dans le putamen et le noyau caudé. Cette hypofixation peut toucher un ou deux hémisphères. La Figure 9.1 issue de [Booth et al., 2015], montre un exemple d'image DatScan dans le cas sain et dans des cas de MP.

Nous avons considéré les données de 59 sujets (40 MP dont 24 hommes et 16 femmes, et 19 sains dont 5 hommes et 14 femmes) de notre base, ayant effectué l'examen DatScan de neuroimagerie, avec un appareil utilisant la technologie hybride CT-SPECT Discovery 670 Pro General Electric.

Les données extraites, par nos collègues de l'hôpital Pitié Salpêtrière, à partir des images des DatScan sont les valeurs des potentiels de liaison dans la partie bilatérale sensorimotrice du putamen (région connue pour être spécialement altérée dans la maladie de Parkinson). Ces potentiels de liaison reflètent la quantité de transporteurs dopaminergiques dans cette région. Les données des DatScan ont été corrigées de manière à prendre en compte l'effet des volumes partiels et segmentées en sous régions striatales anatomiques et fonctionnelles d'après le YeB atlas [Bardinet et al., 2009].

La valeur moyenne de quantité de transporteur dopaminergique est de 5.89 ± 0.65 pour le groupe sain, et de 2.63 ± 0.45 pour le groupe MP. On constate bien une diminution de la quantité de transporteur dopaminergiques chez les MP. Pour connaître la significativité de cette diminution, nous avons effectué une analyse ANOVA à 2 niveaux, en prenant le genre comme un des facteurs. Nous avions préféré cette méthode à celle consistant à faire un Welch t-test

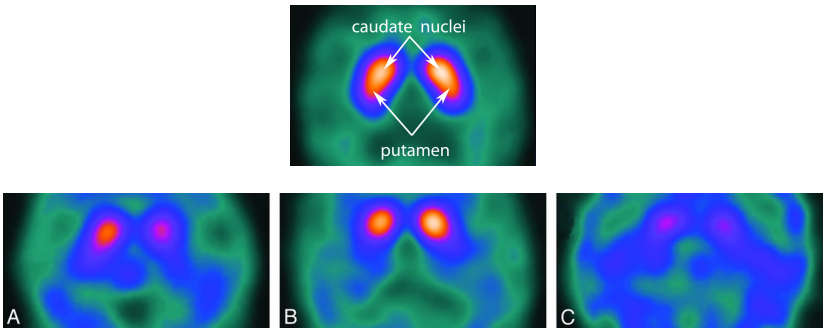


FIGURE 9.1 – Exemple de DatScan d'un sujet sain (en haut) et de sujets MP (en bas). Le DatScan du sujet sain montre une présence symétrique du traceur au niveau du noyau caudé et du putamen, avec une activité de fond très faible. Les DatScan du bas représentent trois cas d'altérations que l'on peut rencontrer chez les MP. L'image A montre une activité asymétrique avec une présence réduite du traceur au niveau du putamen dans un hémisphère. L'image B montre une présence réduite du traceur dans le putamen cette fois dans les deux hémisphères. L'image C révèle une absence de traceur dans le putamen et dans le noyau caudé ainsi qu'une augmentation de l'activité de fond. Source : [Booth et al., 2015]
.

pour les hommes et un pour les femmes, car le nombre de sujets analysés était trop faible pour séparer la base en fonction du genre. Nous avons obtenu $p = 4E - 27$, ce qui montre le grand pouvoir discriminant du DatScan pour séparer les sujets sains des MP.

9.2.2 IRM sensible à la neuromélanine

L'IRM sensible à la neuromélanine (NM) est récemment apparue comme un potentiel biomarqueur prometteur de détection précoce de la maladie de Parkinson. Elle met en évidence une diminution du pigment neuromélanine dans la substance noire compacte [Sulzer et al., 2018] chez les sujets MP, correspondant à la mort des neurones dopaminergiques de la substance noire.

Les données d'IRM sensible à la NM, enregistrées avec un appareil Siemens PRISMA FIT 3 Tesla, ont fait l'objet d'une première analyse sur 117 de nos sujets (85 MP dont 54 hommes et 31 femmes, et 32 sains dont 16 hommes et 16 femmes).

Les volumes de la substance noire basés sur la neuromélanine ont été calculés, par nos collègues de l'Institut du Cerveau et de la Moelle, à partir des images IRM. Les contours de la substance noire ont été dessinés manuellement par deux expérimentateurs, à l'aide du logiciel FreeSurfer5 (http ://freesurfer.net/, Boston, MA, USA) comme cela a été présenté lors d'une étude précédente [Pyatigorskaya et al., 2018]. Les volumes des régions d'intérêt basés sur la neuromélanine ont été calculés à l'aide du logiciel Matlab comme étant le nombre de voxels dans trois images contiguës multiplié par la taille d'un voxel. Le volume de la substance noire a ensuite été divisé par le volume total intracranial pour normaliser par rapport à la taille de la tête.

Le volume moyen de la substance noire est de $2.7E-4 \pm 4.96E-5$ pour le groupe sain, et de $2.2E-4 \pm 5.0E-5$ pour le groupe MP. On constate comme attendu une diminution du volume de la substance noire chez les MP. Pour connaitre la significativité de cette diminution, nous de nouveau avons effectué une ANOVA à 2 niveaux. Nous avons obtenu $p = 3E - 6$, ce qui montre un pouvoir discriminant mais moins élevé que celui du DatScan ($p = 4E - 27$) pour séparer les sujets sains des MP.

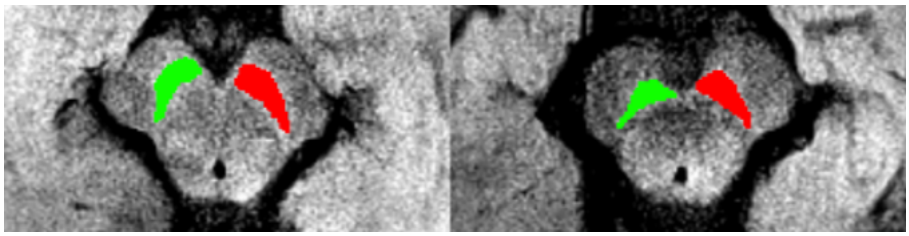


FIGURE 9.2 – Images d'IRM sensible à la neuromélanine, sur deux de nos sujets, avec à gauche un sujet sain et à droite un sujet MP. Nous pouvons constater une diminution de la substance noire (parties colorées) chez le sujet MP. Source : [Gaurav et al., 2019]

9.2.3 Scores moteurs

L'échelle MDS-UPDRS [Goetz et al., 2007], présentée partie 1.1.4, est l'échelle d'évaluation de référence pour quantifier la progression de la maladie de Parkinson. Elle est composée de quatre parties dont la partie III qui consiste en un examen moteur effectué, par nos collègues de l'hôpital Pitié Salpêtrière, chez tous les sujets de la base ICEBERG en OFF et en ON (pour ceux qui prennent un traitement). Nous avons analysé les données de cette partie III (effectuée en OFF) car c'est la partie la plus adaptée à l'évaluation analytique de la motricité dans la maladie de Parkinson. Le score moteur moyen est de 34.06 ± 7.15 pour le groupe MP et de 4.19 ± 3.56 pour le groupe sain. On constate comme prévu un score moteur plus élevé pour les MP. Pour connaitre la significativité de cet écart, nous de nouveau avons effectué une ANOVA à 2 niveaux. Nous avons obtenu $p = 3E - 46$, ce qui montre un pouvoir discriminant élevé pour séparer les sujets sains des MP.

9.3 Paramètres vocaux

Nous avions extrait 19 des paramètres vocaux globaux présentés partie 7.1, à partir des enregistrements issus du microphone professionnel.

La phonation était caractérisée par 13 paramètres extraits lors de la tâche de voyelle soutenue (DUV, DVB, jitter local, jitter rap, jitter ppq5, jitter ddp, shimmer local, shimmer apq3, shimmer apq5, shimmer apq11, shimmer dda, NHR, HNR).

La prosodie par *SD log Fo* extrait lors de la lecture et du monologue.

Enfin la répartition des pauses du monologue était décrite par :
- le nombre de pauses $<$ 200ms ;
- le nombre de pauses d'une durée $\in$ [200, 500ms] ;
- le nombre de pauses $\in$ [500ms, 1s] ;
- le nombre de pauses $>$ 1s ;
- la médiane des pauses $>$ 200ms.

Pour savoir quels paramètres étaient significativement discriminants entre les groupes MP et sains, et savoir quels paramètres garder pour les corrélations, nous avions effectué une analyse de variance (ANOVA) à deux niveaux, avec le sexe des sujets comme un des deux niveaux, pour prendre en compte la sensibilité de ces paramètres vocaux au genre. Suite à cette analyse nous avons retiré des paramètres redondants et garder les paramètres les plus discriminants. Les résultats de l'ANOVA des 7 paramètres gardés pour les corrélations sont présentés Figure 9.3. Parmi ces résultats, nous retrouvons la diminution significative du paramètre prosodique *SD log*

Fo chez les MP, ainsi que la diminution du nombre de pauses courtes et l'augmentation de la durée médiane des pauses.

Paramètres vocaux	Tâches	MP homme	Sain homme	MP femme	Sain femme	p (MP vs Sain)
Phonation						
DVB	voyelle soutenue	1,27E-02	8,04E-03	2,40E-02	2,69E-02	0,95
jitter local	voyelle soutenue	9,87E-03	8,17E-03	8,32E-03	6,99E-03	0,45
shimmer local	voyelle soutenue	5,46E-02	5,79E-02	3,88E-02	3,20E-02	0,73
HNR	voyelle soutenue	19,66	18,57	23,85	23,52	0,40
Prosodie						
SD log fo	lecture+monologue	5,64E-02	7,27E-02	6,83E-02	7,49E-02	**2,70E-06**
Pauses						
Nb pauses 200 à 500ms	monologue	0,22	0,32	0,28	0,39	**0,0062**
Médiane pauses > 200ms	monologue	0,74	0,60	0,53	0,48	**0,041**

FIGURE 9.3 – Moyennes des paramètres vocaux retenus pour les corrélations chez les patients MP hommes, MP femmes et chez les sujets sains hommes et femmes. Les valeurs-p entre les MP et les sains ont été calculées à partir d'une ANOVA à 2 niveaux, prenant en compte les différences vocales dues au genre. DUV=Degree of Unvoiced Segments, DVB=Degree of Vocal Breaks, NHR=Noise-to-harmonics ratio, HNR=Harmonics-to-Noise Ratio, fo=Fréquence fondamentale, MP=Maladie de Parkinson

9.4 Corrélations

De manière à étudier le lien entre les paramètres vocaux et les données cliniques et d'imagerie, nous avons construit un modèle de régression linéaire multiple, cf. équation 9.1, auquel nous avons appliqué comme variables prédictives (X_i), supposées indépendantes, notre ensemble de paramètres vocaux sélectionné précédemment et le genre. Les variables réponses que l'on cherche à expliquer (Y_i) sont successivement : le score d'UPDRS III OFF, la quantité de transporteurs dopaminergiques (DatScan) et le volume de la substance noire (IRM sensible à la NM). Les indices i correspondent aux différents sujets, n étant le nombre de sujets. ϵ_i est l'erreur du modèle qui résume l'information manquante dans l'explication linéaire des valeurs de Y_i à partir des $X_{i1}, ..., X_{ip}$. On suppose la distribution de ϵ_i gaussienne et de moyenne nulle. Les coefficients $a_0, a_1, ..., a_p$ sont les paramètres à estimer, p étant le nombre de variables prédictives (soit $p = 8$ dans notre cas).

$$Y_i = a_0 + a_1 X_{i1} + a_2 X_{i2} + ... + a_p X_{ip} + \epsilon_i \qquad \text{avec i=1, ..,n} \tag{9.1}$$

L'objectif est d'expliquer les données Y d'imagerie et d'UPDRS III par une combinaison linéaire de nos 7 paramètres vocaux X_k et l'information de genre. Les coefficients a_k sont estimés de manière à en minimiser l'erreur de prédiction par une méthode de moindre carré. Plus précisément ils sont estimés de manière à ce que la moyenne de l'erreur de prédiction sur l'ensemble des sujets soit nulle, et sont optimisés pour réduire au maximum la variance de cette erreur de prédiction. Chacun de ces coefficients a_k exprime l'impact d'une variation d'une unité de la variable prédictive X_k sur la moyenne de la variable réponse Y, son signe traduisant le sens de l'effet. Cependant un coefficient plus élevé qu'un autre ne signifie pas forcément que la variable prédictive correspondante a plus d'impact sur la variable réponse, car les variables prédictives n'ont pas forcément les mêmes ordres de grandeur. Pour estimer l'effet d'une variable prédictive sur la variable réponse, il faut diviser l'estimation de son coefficient (*Estimate*) par l'erreur standard associée (*SE*). Le résultat obtenu est ce qu'on appelle la statistique-t (tStat) (cf. équation 9.2) correspondant au test de l'hypothèse nulle selon laquelle la variable prédictive n'aurait pas d'effet sur la variable résultat (donc le coefficient associé serait nul), sachant les autres variables prédictives du modèle.

$$tStat = \frac{Estimate}{SE} \tag{9.2}$$

Enfin la valeur-p correspondante à la statistique-t est calculée afin de donner le niveau de significativité correspondant. Les estimations des coefficients, l'écart standard associé, les statistiques-t et la valeurs-p sont détaillées Figure 9.4.

	Estimate	SE	tStat	pValue
(Intercept)	9.61	16.39	0.59	0.56
DVB	12.49	21.66	0.58	0.57
jitter local	-42.55	192.14	-0.22	0.83
shimmer local	180.88	95.34	1.90	0.06
HNR	1.12	0.53	2.11	0.04
SD log fo	-302.54	88.58	-3.42	0.00
nb pauses 200 à 500ms	-0.58	6.13	-0.09	0.92
médiane pauses > 200ms	7.46	4.92	1.52	0.13
sexe	-4.24	2.89	-1.47	0.14

	Estimate	SE	tStat	pValue
(Intercept)	5.85	2.76	2.12	0.04
DVB	6.66	4.61	1.45	0.15
jitter local	-79.12	44.12	-1.79	0.08
shimmer local	-20.81	16.20	-1.28	0.20
HNR	-0.20	0.10	-2.14	0.04
SD log fo	42.48	13.01	3.27	0.00
nb pauses 200 à 500ms	1.48	0.79	1.87	0.07
médiane pauses > 200ms	0.02	0.92	0.02	0.98
sexe	0.93	0.47	1.98	0.05

	Estimate	SE	tStat	pValue
(Intercept)	0.00	0.00	3.50	0.00
DVB	-0.00	0.00	-0.25	0.81
jitter local	-0.00	0.00	-0.16	0.87
shimmer local	-0.00	0.00	-1.16	0.25
HNR	-0.00	0.00	-1.04	0.30
SD log fo	0.00	0.00	1.93	0.06
nb pauses 200 à 500ms	-0.00	0.00	-0.07	0.94
médiane pauses > 200ms	-0.00	0.00	-0.75	0.46
sexe	0.00	0.00	2.36	0.02

FIGURE 9.4 – Estimation des coefficients (*Estimate*) des modèles de régression linéaire multiple, avec le score UPDRS III (en haut), le DatScan (au milieu) et le volume de la substance noire (en bas) comme variables réponses, et 7 paramètres vocaux et l'information du genre comme variables prédictives. Les écarts standards associés (SE), les statistiques-t (tStat) et la valeurs-p sont également précisées, ainsi que le coefficient a_0 (*Intercept*)

Nous pouvons constater que SD log Fo est le paramètre vocal expliquant le mieux les variables réponses UPDRS III, et les variables d'imagerie, avec $p < 0.01$ pour l'UPDRS III et le DatScan. La variable HNR a également un effet significatif sur l'UPDRS III et les données du DatScan. Si on observe l'ensemble des paramètres, on constate qu'ils semblent mieux expliquer les valeurs d'UPDRS III et de DatScan que le volume de la substance noire estimé à partir des images d'IRM sensible à la NM, ce qui peut venir du fait que le volume de la substance noire tel qu'il a été calculé est moins discriminant que l'UPDRS et les données du DatScan, cf. section 9.2.2.

Concernant les statistiques des deux paramètres liés aux pauses, on observe pour l'analyse avec l'UPDRS et celle avec le DatScan que chaque fois une de ces deux variables prédictives a un très faible poids statistique, ce qui peut indiquer une probable redondance entre ces deux paramètres.

Les statistiques concernant la prédiction du modèle linéaire, composé de l'ensemble des 7 paramètres vocaux et de l'information de genre, sont présentées au Tableau 9.1.

- On y trouve la *Root Mean Squared Error* (RMSE) définie comme la racine de l'erreur quadratique moyenne. La RMSE est égale à l'écart type des estimations de la variable résultats si l'estimateur est bien non biaisé (c'est-à-dire erreur de prédiction de moyenne nulle). Pour une même variable résultat, plus la RMSE est grande, moins le modèle de régression linaire est précis. Si on compare les RMSE de différentes variables résultats, il faut prendre en compte les différences d'ordre de grandeur de ces dernières.

- Le paramètre *R2*, aussi appelé coefficient de détermination, est plus adapté pour comparer les précisions de différents modèles de régressions appliqués à différentes variables résultats. Ce paramètre indique la proportion de la variance d'une variable résultat expliquée par le modèle de régression linéaire. Par exemple ici notre modèle de régression linéaire explique 42% de la variance des scores de DatScan, 21% de celle des scores UPDRS et 19% de celle des données d'IRM sensible à la NM.

- Les statistiques F sont aussi précisées, elles représentent le résultat du test F qui compare la précision de notre modèle de régression linéaire à celle qu'on aurait avec un modèle composé seulement d'une constante.

- Enfin la valeur-p indique la significativité liée au test F du modèle.

Variables résultats	Examen	RMSE	R2	F	valeur-p
UPDRS III OFF	tests moteurs	0.215	0.215	4.42	9.56E-5
Transporteurs dopaminergiques	DatScan	1.34	0.422	4.47	4.06E-4
Volume substance noire	IRM NM	5.2E-5	0.191	3.11	3.32E-3

TABLE 9.1 – Résultats statistiques des régressions linéaires multiples pour les trois types de données cliniques et d'imagerie que l'on a cherché à prédire à partir des 7 paramètres vocaux (DVB, jitter local, shimmer local, HNR, SD log Fo, nombre des pauses entre 200 à 500ms et durée médiane des pauses) et de l'information du genre.

On constate que pour les trois variables résultats, qui sont le score UPDRS III, la quantité de transporteurs dopaminergiques et le volume de la substance noire, les modèles de régression linéaires ont des performances significatives (p allant de 3.32E-3 à 9.56E-5). Ce qui signifie que les 7 paramètres vocaux et l'information de genre sont capables de prédire linéairement de manière significative les résultats moteurs, et les données de DatScan et d'IRM sensible à la NM. Ces paramètres vocaux corrèlent particulièrement avec les données du DatScan car ils permettent d'expliquer 42% de la variance de la quantité de transporteur dopaminergique.

9.5 Conclusion sur corrélations voix avec neuroimagerie et paramètres cliniques

En conclusion à partir de 7 paramètres vocaux extraits sur une sous partie de notre base de données, lors de la lecture, du monologue et des voyelles soutenues enregistrés avec le microphone professionnel, caractérisant la prosodie, la répartition des pauses et la phonation, nous avons pu expliquer, grâce à un modèle de régression linéaire multiple, 42% de la variance des données du DatScan, 19% de celle des données d'IRM sensible à la neuromélanine et 21% de celle des scores UPDRS III. C'est à dire que ces 7 paramètres, ainsi que l'information du genre, sont capables de prédire linéairement ces données de manière significative.

La précision concernant la prédiction des données du DatScan est intéressante car ces données caractérisent particulièrement bien l'évolution de MP, notamment au stade débutant. Les seuls inconvénients de cet examen est son coût élevé et son accessibilité réduite. L'intérêt de poursuivre l'analyse de corrélations avec la voix serait de pouvoir trouver un ensemble optimal de

paramètres vocaux qui permettrait de prédire la quasi-totalité de la quantité de transporteurs dopaminergiques, et donc d'avoir un examen moins coûteux et plus accessible pouvant fournir des informations équivalentes, et aider au diagnostic et au suivi de l'évolution de la maladie. Plusieurs améliorations peuvent être apportées à l'ensemble de paramètres vocaux qu'on a utilisé, comme la suppression d'un des deux paramètres liés aux pauses (pour enlever une redondance), et l'ajout d'autres paramètres liés par exemple à l'articulation ou à la capacité de suivre un rythme constant.

Dans cette analyse nous nous sommes intéressés, en ce qui concerne le DatScan, seulement aux corrélations avec la région du cerveau la plus discriminante dans la MP, à savoir la partie bilatérale sensorimotrice du putamen. Les sous parties limbiques et associatives du putamen ont aussi été analysées, ainsi que d'autres parties du striatum, comme le noyau caudé et le noyau accumbens, segmentés également en sous régions. Il serait intéressant de tester le pouvoir prédictif de nos paramètres vocaux sur l'ensemble de ces régions, et d'étudier quel type de paramètre corrèle avec quelle région. Comme les paramètres vocaux reflètent chacun certains troubles spécifiques liés à la maladie de Parkinson, cela pourrait nous permettre de mieux comprendre les différentes altérations des circuits neuronaux dans les premiers stades de la maladie de Parkinson. Également, faire l'analyse sur plus de sujets nous permettra de séparer les hommes des femmes, dans les modèles de régression linéaire, et de comprendre ces altérations genre par genre, afin de mieux comprendre les éventuelles différences de mécanisme d'altération de ces réseaux entre les hommes et les femmes, cf. [Haaxma et al., 2007].

Enfin les corrélations des paramètres vocaux avec d'autres variables comme des scores cognitifs ou des scores génétiques pourraient également s'avérer utiles, afin de pouvoir prévoir, par exemple, certains déclins cognitifs associés à la maladie de Parkinson.

Chapitre 10

Conclusion générale

Pour résumer nous nous sommes intéressés à la détection automatique de MP au stade débutant à partir de l'analyse de la voix. Pour cela nous avons commencé par constituer des bases de données voix de plus de 200 sujets, comprenant des sujets MP débutants (dont le diagnostic remontait à moins de 4 ans), des sujets sains et des sujets iRBD, considérés au stade prodromique de la maladie de Parkinson. Ces sujets ont été enregistrés pendant une quinzaine de minutes avec un microphone professionnel, et en simultané avec le microphone interne d'un ordinateur. Ils ont également effectué une fois par mois des enregistrements vocaux, en appelant un serveur vocal interactif, à partir de leur propre téléphone. Au cours de ces enregistrements, les sujets ont effectué différentes tâches vocales, comme des voyelles soutenues, des répétitions de phrases, de la lecture, des répétitions rapides de syllabes (DDK), des répétitions lentes de syllabes à un rythme imposé, et un monologue au cours duquel ils ont raconté leur journée.

Nous avons analysé ces enregistrements par le biais de 3 méthodes d'analyses différentes, faisant intervenir différentes échelles de temps, différents paramètres vocaux (liés à différents domaines phonétiques), et différents classifieurs. Les deux premières méthodes (cf. chapitres 5 et 6) sont inspirées des méthodes utilisées en reconnaissance du locuteur. Elles utilisent toutes les deux des paramètres cepstraux, les MFCC, caractérisant l'enveloppe spectrale, donc plutôt liés à l'articulation. Pour la méthode MFCC-GMM, la classification s'opère à l'échelle de la trame (fenêtre de 20ms), alors que pour la méthode des x-vecteurs, la classification se fait au niveau du segment (3s). La troisième méthode, utilisée dans le chapitre 7, utilise des paramètres dits globaux, calculés à l'échelle des tâches, et reflétant d'autres domaines de la voix, comme la prosodie, la phonation, la fluence verbale, et la capacité à suivre un rythme imposé. Enfin une fusion de ces trois méthodes a été effectuée. Nous avons fait toutes les analyses en traitant séparément les hommes des femmes, afin de ne pas rajouter la variabilité due au genre, et afin d'évaluer d'éventuelles différences, selon le genre, dans les changements vocaux dus à MP.

La première étape de toutes ses analyses a consisté en divers prétraitements (comme la soustraction spectrale), afin entre autres de supprimer l'effet du non appariement complet de l'environnement acoustique entre les groupes, dû aux différents lieux d'enregistrement.

Analyse MFCC-GMM

La première méthode de classification que nous avons choisi d'utiliser est celle des MFCC-GMM, car elle avait l'avantage de nécessiter peu de données et d'avoir un faible coût computationnel. Nous avons entrainé des modèles GMM pour décrire la distribution des MFCC de sujets MP et de sujets sains d'entraînement, et utilisé la log-vraissemblance (LLH) pour tester les vecteurs MFCC de sujets tests, par rapport au modèle MP et au modèle sain. Nous avons ensuite calculé un score compris entre 0 et 1 à partir des ratios des LLH, moyennées sur l'ensemble des

trames testées.

Les meilleures performances ont été obtenues à partir des tâches de lecture et répétitions de phrases ainsi qu'à partir des tâches DDK (la tâche /pataka/ étant la plus performante). La tâche de type monologue s'est révélée un peu moins efficace, ce qui peut s'expliquer par la variabilité de son contenu phonétique d'un sujet à l'autre, inhérente aux tâches texte-indépendant, pouvant masquer une partie de la variabilité due à MP. Enfin les répétitions lentes et surtout les voyelles soutenues se sont révélées peu appropriées pour ce type d'analyse.

Nous avons également évalué l'influence du contenu des données utilisées pour l'entraînement au regard des tâches utilisées pour le test. Nous avons constaté que le choix optimal des données utilisées pour l'entraînement des GMM résulte d'un compromis entre quantité et spécificité.

La fusion des deux meilleures tâches, à savoir la lecture + répétition de phrase testées par rapport à des GMM spécifiques, et la tâche /pataka/ testée par rapport à des GMM globaux, a conduit à un EER de 17% (Acc=83%) chez les hommes, à partir de 1min30 de paroles par sujet test, enregistrées avec le microphone professionnel.

Avec le microphone de l'ordinateur, nous avons observé une dégradation moyenne de 8%, lors de la classification des hommes MP vs sain. Cette dégradation étant légèrement moins importante pour la lecture mais plus importante pour les tâches DDK. Les causes de cette dégradation sont liées à la distance accrue entre la bouche et le microphone et à la fonction de débruitage actif de l'ordinateur. La qualité réduite du microphone de l'ordinateur pouvant également contribuer légèrement à cette dégradation.

Concernant les enregistrements téléphoniques, nous avons constaté une dégradation supplémentaire des performances, pour la classification des hommes MP vs sain, quand on utilisait une session téléphonique par sujet test. Une simulation simple du téléphone, à partir de nos enregistrements issus du microphone professionnel, nous a permis de comprendre que cette dégradation résultait pour moitié de l'échantillonnage plus faible et de la bande de fréquence étroite. L'autre moitié serait la conséquence des autres caractéristiques du téléphone, comme le bruit, la distorsion due aux codecs.. ainsi que l'exécution non supervisée des tâches.

Nous avons également constaté une amélioration des performances de 10% en prenant plus de données parole par sujet test (considérant toutes les sessions téléphoniques pour le test). Ceci aboutit à un EER de 25% (Acc=75%) pour la détection des hommes MP vs sain, avec une moyenne de 5min de parole DDK par sujet test.

Nous avons effectué une analyse complémentaire en ciblant la classification MFCC-GMM à l'attaque des sons voisés. Nous avons constaté que les attaques des occlusives /p/ étaient spécialement discriminantes dans la maladie de Parkinson, car une classification à partir uniquement de ces sons a conduit à un EER de 27% (avec seulement l'équivalent de 2s de données par sujet testé).

Pour conclure sur cette analyse, la méthode de classification MFCC-GMM s'est avérée pertinente pour la détection des hommes MP débutants, avec un EER de **17%** pour les enregistrements du microphone professionnel, et **25%** pour les enregistrements téléphoniques. Pour la détection de MP débutant chez les femmes, cette méthode ne s'est pas révélée efficace (avec des EER autour de 40%), ce qui peut être dû, entre autres, à la plus grande variabilité des MFCC chez les femmes.

Analyse à partir des x-vecteurs

Dans le chapitre 6, nous avons adapté la dernière méthode en date utilisée en reconnaissance du locuteur, dont les performances dépassent celles des GMM dans ce domaine, mais nécessitant

beaucoup de données et étant plus coûteuse computationnellement. C'est la première fois que cette méthode est utilisée dans le cadre de la détection de MP.

Cette méthode se base sur l'extraction d'*embeddings*, appelés x-vecteurs, extraits à partir d'un DNN prenant en entrée des vecteurs MFCC. Nous avons fait varier différentes conditions, tout en comparant, pour chaque condition, 3 méthodes de classification (distance cosinus, LDA + distance cosinus et PLDA). Comme l'entraînement du DNN nécessite généralement beaucoup de données, nous avons utilisé un DNN pré-entraîné pour la reconnaissance du locuteur.

Les analyses sur notre base téléphonique concernant la classification des hommes MP vs sains, nous ont permis de constater que les performances étaient meilleures quand les segments audio testés avaient la même durée (3s) que les segments ayant servi pour l'entraînement (du DNN, de la LDA et de la PLDA) et pour la constitution des x-vecteurs moyens MP et sain.

Concernant la comparaison des 3 types de classifications, on constate dans l'ensemble une nette amélioration des performances quand on ajoute une LDA avant le calcul de distance cosinus. On constate également une performance équivalente entre LDA + distance cosinus et la PLDA.

Nous avons également constaté qu'effectuer une augmentation de données (en dupliquant nos données avec rajout de divers bruits) améliore les performances du monologue. Cela n'améliore, par contre, pas les performances des tâches plus texte-dépendant, ce qui est cohérent avec le fait que l'augmentation de données, en rajoutant du bruit de différentes sortes, nuit à la spécificité du contenu phonétique.

Si on compare les performances avec celles de notre classifieur MFCC-GMM, nous pouvons constater une amélioration des performances de classification pour la tâche de monologue. Ce qui est cohérent avec le fait que les x-vecteurs ont été à l'origine élaborés pour la reconnaissance du locuteur indépendante du texte.

Les tâches très spécifiques, comme les DDK, présentent quant à elles, de meilleurs performances avec les GMM qu'avec les x-vecteurs. Ceci pouvant être la conséquence du DNN pré-entraîné pour la reconnaissance du locuteur à partir de données paroles beaucoup plus variées que les phonèmes prononcés lors les tâches DDK.

Dans le but de rendre le DNN plus spécifique aux tâches DDK, nous avons effectué une analyse complémentaire en l'entrainant cette fois avec notre base de données (à partir des tâches DDK). Les performances obtenues n'ont pas montré d'amélioration par rapport au DNN pré-entraîné pour la reconnaissance du locuteur. Ceci pouvant être dû à la quantité réduite de nos données disponibles pour l'entraînement du DNN (nécessitant habituellement beaucoup de données).

Enfin le dernier résultat à souligner est la nette amélioration des performances, par rapport à la méthode MFCC-GMM, pour la détection de MP chez les femmes. L'EER est réduit d'environ 10% pour le monologue (7% à partir des enregistrements téléphoniques et 15% à partir du microphone professionnel). Cette amélioration pourrait provenir de l'apport de la LDA, dont le principe est de diminuer la variabilité intraclasse, en augmentant la variabilité interclasses. Ainsi avec la classification x-vecteur combinée à une LDA et une distance cosinus, nous arrivons à détecter les femmes MP débutants avec un EER de **30%** à partir du microphone professionnel (avec environ 1 min de parole par sujet test) et de **33%** à partir des enregistrements téléphoniques (avec environ 5 min de parole par sujet test).

Ces deux types de classification (MFCC-GMM et x-vecteur) permettent une détection de la maladie de Parkinson au stade débutant avec une précision (Acc) de 83% pour les hommes et

70% pour les femmes (avec le microphone professionnel) aux seuils EER, en exploitant quasiment uniquement les troubles articulatoires. Or les altérations vocales rencontrées dans la maladie de Parkinson ne concernent pas seulement l'articulation, mais aussi la prosodie, la phonation, le débit de parole et les habilités rythmiques. Nous avons donc voulu analyser également ces autres domaines afin d'enrichir les informations vocales dont nous pouvons disposer pour détecter MP précocement.

Analyse des paramètres globaux

Nous avons ainsi extrait des paramètres attraits à la prosodie, à la phonation, à l'utilisation des pauses, et à la capacité à suivre un rythme. Ces paramètres sont dits globaux, car ils sont calculés à l'échelle de la tâche vocale.

Nous avons effectué des analyses de variance afin de savoir quels paramètres différaient de manière significative entre les groupes MP et sains, et quelles tâches vocales mettaient le mieux en valeur ces différences. Les paramètres qui se sont révélés les plus discriminants sont :

- *SD log Fo*, dont la diminution chez les MP traduit la monotonie de l'intonation. La diminution de la prosodie concerne les répétitions de phrases, le monologue, et particulièrement la lecture du dialogue à contenance émotionnelle.
- Le nombre de pauses $\in$ [200ms, 500ms] (réduit chez les MP) et la médiane des pauses $>$ 200ms (allongée chez les MP) extraits du monologue. Les altérations de ces paramètres traduisent un débit de parole saccadé chez les MP.
- La variation relative du rythme (*RSD*) lors de la répétition lente de syllabes et l'*Ecart* moyen avec le rythme imposé. Ces paramètres sont augmentés chez les MP, traduisant une difficulté à garder un rythme constant et suivre un rythme imposé. Ces variations rythmiques sont majorées lors de la répétition alternée des syllabes /pa/ et /kou/, par rapport aux répétitions non alternées.

Pour tous ces paramètres on constate des différences entre les groupes MP et sain, que ce soit chez les hommes ou chez les femmes. Néanmoins les différences sont plus marquées chez les hommes.

Les variations de l'intensité et les paramètres relatifs à la phonation, se sont révélés peu discriminants. Ceci peut s'expliquer par le fait que les altérations de ces deux types de paramètres sont connues pour être atténuées par les traitements dopaminergiques [Rusz et al., 2013b], or quasiment tous nos patients sont traités et ont été enregistrés en ON.

A partir de ces analyses nous avons choisi un ensemble réduit de 6 paramètres pour effectuer une classification, MP vs sain, de type SVM, avec une fonction noyau linéaire. L'ensemble de paramètres comprend : SD log Fo extrait dans deux conditions différentes, les deux paramètres liés aux pauses cités précédemment, et les deux paramètres rythmiques extraits lors des tâches /pa kou/. Nous avons obtenu comme résultat, que ce soit avec le microphone professionnel ou avec le microphone de l'ordinateur, un EER de **22%** pour les hommes, et d'environ **31%** pour les femmes, le tout à partir de 3 min de parole par sujet.

Concernant les enregistrements téléphoniques, les meilleurs résultats ont été obtenus, en considérant seulement le paramètre prosodique *SD log Fo*. Les EER correspondants sont de 27% pour les hommes et 33% pour les femmes, avec environ 6 min de données parole par sujet. Ces performances de classification sont légèrement inférieures aux performances obtenues avec les deux analyses précédentes utilisant les MFCC. Néanmoins cette analyse reste intéressante car elle exploite des caractéristiques différentes de la voix, et est donc porteuse d'informations

complémentaires concernant les altérations vocales dans la maladie de Parkinson.

Fusion

Afin de prendre en compte les différentes informations, quant au caractère parkinsonien d'une voix, issues de ces méthodes d'analyses, nous avons effectué une fusion de ces trois méthodes. Nous avons opté pour une méthode simple de fusion à vote majoritaire. Nous avons constaté une amélioration de 6% par rapport au meilleur classifieur, pour la détection des hommes MP enregistrés avec le microphone professionnel. La performance de classification s'élevant alors à **Acc=89% (avec Sp=92% et Se=87%)**.

Concernant les enregistrements téléphoniques, la fusion n'a pas amélioré les résultats. Ce qui est probablement dû aux moins bonnes performances des trois classifieurs, en effet plus les performances sont élevées plus la fusion a de chances d'améliorer les résultats.

Enfin pour les femmes, ce type de fusion n'a pas non plus amélioré les performances de classification. Ceci résulterait de la mauvaise performance de la méthode MFCC-GMM pour les femmes, qui aurait un impact négatif sur la décision par vote majoritaire.

Effet du genre

A partir de ces trois méthodes de classification, nous avons constaté un gros effet de genre, avec de moins bonnes performances de classification pour les femmes. Plusieurs raisons sont à l'origine de ces différences. Tout d'abord, la plus grande variabilité des MFCC chez les femmes [Fraile et al., 2009b] semble nuire considérablement à la détection de MP par la méthode MFCC-GMM. Nous avons cependant pu réduire cette variabilité avec l'utilisation de la LDA sur les x-vecteurs, permettant une classification des femmes avec un EER de 30% (avec le microphone professionnel), pour un EER de 25% pour les hommes dans les mêmes conditions. Cette diminution des performances de classification chez les femmes est également rencontrée dans l'analyse à partir des paramètres globaux.

Ces moins bonnes performances dans la détection de la MP débutante chez les femmes, pourrait s'expliquer en partie par une atteinte neuronale en moyenne moins marquée chez les femmes MP que chez les hommes MP [Haaxma et al., 2007] et une symptomatologie plus bénigne chez les femmes, d'après cette même étude. De plus une apparition des symptômes de MP plus tardive de 2 ans en moyenne a été observée chez les femmes, comparé aux hommes. L'effet éventuellement protecteur de l'œstrogène sur la maladie de Parkinson a souvent été avancé pour expliquer les différences dans l'expression de la MP selon le genre.

Nous pouvons d'ailleurs constater dans notre base de données que le score moteur de l'UPDRS III des sujets MP est en moyenne plus élevé chez les hommes que chez les femmes (34 pour les hommes et 29 pour les femmes). Ce qui peut aussi contribuer à expliquer la plus grande difficulté à détecter MP chez les femmes de notre base de données.

Enfin des études ont montré que les circuits neuronaux de la parole étaient différents chez les hommes et chez les femmes [de Lima Xavier et al., 2019, Jung et al., 2019]. Ces circuits neuronaux peuvent donc être différemment impactés par la MP, et conduire à différents types ou différents degrés d'altérations vocales suivant le genre.

Influence des méthodes ensemblistes

Des méthodes ensemblistes ont été utilisées pour les différentes classifications, par agrégation des scores d'un *repeated random subsampling* pour les analyses à partir des MFCC, et par agrégation des scores d'un 10 fois 10-fold pour les classifications à partir des paramètres glo-

baux. Nous avons comparé les performances obtenues à celles des classifieurs non agrégés et avons constaté une amélioration de 2 à 3% pour les analyses MFCC-GMM et x-vecteurs. Nous n'avons pas trouvé d'amélioration avec l'agrégation pour l'analyse des paramètres globaux. Ceci semble indiquer que notre classifieur SVM est suffisamment stable et ne nécessite du coup pas d'agrégation, alors qu'elle semble plutôt bénéfique aux classifeurs MFCC-GMM et x-vecteur.

Influence du type de tâches vocales

Pour chacune des 3 méthodes de classifications, nous avons évalué la pertinence des différentes tâches vocales au regard du type d'analyse effectuée. Les tâches vocales conduisant aux meilleures performances pour la classification MFCC-GMM sont les tâches texte-dépendant, à savoir la lecture, les répétitions de phrases et les tâches DDK (surtout la tâche /pataka/). En effet la variabilité du contenu phonétique inter-sujets étant réduite, la variabilité due à MP est plus facilement détectable. Ceci ne s'applique néanmoins pas aux voyelles soutenues. Pour ces dernières, le contenu phonétique restreint au phonème /a/ ne semble pas contenir assez d'informations discriminantes pour détecter MP au stade débutant.

Concernant les classifications à partir des x-vecteurs + LDA et distance cosinus, les tâches les plus appropriées semblent le monologue, la lecture et les répétitions de phrases. Les tâches plus spécifiques, comme les DDK, sont sous-exploitées par ce classifieur, car le DNN a été pré-entraîné pour la reconnaissance du locuteur, à partir de données paroles de type conversation, donc au contenu phonétique plus large.

Enfin pour les analyses à partir des paramètres globaux, le monologue et la lecture à contenu émotionnel se sont révélées particulièrement appropriées pour détecter la diminution prosodique chez les MP. Les tâches de lecture et du monologue se sont aussi avérées pertinentes pour la caractérisation de l'altération du débit. Les tâches de répétitions lentes de syllabes ont permis de bien mettre en évidence la diminution des capacités chez les MP à suivre un rythme constant. Cette altération étant majorée quand il s'agit d'alterner les syllabes à répéter (tâche /pa kou/). Enfin les voyelles soutenues montrent une légère altération du timbre, mais pas assez discriminante pour être utilisée lors d'une classification (effet probable de l'amélioration de la phonation par les traitements médicamenteux).

Influence du type de microphone utilisé

Nous avons effectué les différentes classifications à partir des enregistrements issus du microphone professionnel, issus du microphone interne de l'ordinateur et à partir des enregistrements téléphoniques, afin d'étudier l'effet du type d'enregistrements sur les performances de classification.

Nous avons pu constater pour les analyses MFCC-GMM de meilleures performances avec le microphone professionnel qu'avec le microphone de l'ordinateur. On observe en effet une dégradation de 4% avec le microphone de l'ordinateur, à partir des tâches de lecture et répétitions de phrases, et une dégradation de 13% à partir des tâches DDK. Les enregistrements du microphone de l'ordinateur semblent ainsi contenir moins d'informations discriminantes liée à la production des occlusives. Ceci pouvant être la conséquence de la distance accrue entre la bouche et le microphone, et/ou la conséquence de la fonction de débruitage actif de l'ordinateur. La qualité intrinsèques des deux types de microphones peut jouer aussi mais ne peut pas expliquer à elle seule un tel écart. Ces différentes caractéristiques propres aux enregistrements effectués avec le microphone de l'ordinateur ne semblent, par contre, pas avoir d'impact significatif sur les analyses à partir des paramètres globaux.

Nous avons également testé l'effet "cross-micro" pour estimer la performance qu'on pourrait avoir en utilisant nos modèles MFCC-GMM construits à partir des enregistrements du microphone professionnel et en testant des sujets enregistrés dans d'autres conditions avec un autre microphone. Pour cela nous avons utilisé les enregistrements du microphone de l'ordinateur pour le test, et ceux du microphone professionnel pour l'entraînement. Nous avons obtenu un EER de 25% pour la classification des hommes à partir des tâches de lecture et répétition de phrases, ce qui est du même ordre que lorsque tout est fait à partir du microphone de l'ordinateur.

Concernant les enregistrements téléphoniques, ils ont conduit, comme attendu, à de moins bonnes performances que les enregistrements du microphone professionnel, pour les trois types de classification. Pour les analyses MFCC-GMM nous avons montré que cette dégradation des performances semblait provenir pour moitié de limitation de la bande de fréquence et de la fréquence d'échantillonnage réduite, et pour l'autre moitié d'autres causes comme le bruit de fond, la distorsion des codecs et le moins bon respect des consignes dû à l'exécution des tâches en autonomie. Ce dernier aspect semble également avoir son importance dans l'analyse de certains paramètres globaux, comme les paramètres rythmiques. Une VAD moins efficace pour les enregistrements téléphoniques semblerait aussi contribuer à cette diminution de performances. Néanmoins il convient d'interpréter les différences de performance, entre les enregistrements professionnels et téléphoniques, avec précaution car le nombre de sujets, la quantité de données vocales par sujet et les tâches vocales effectuées ne sont pas complètement identiques.

Cas des iRBD

Nous avons également tenté de détecter les iRBD, considérés comme au stade prodromique de MP (ou d'un autre syndrome parkinsonien), par rapport aux sujets sains, en utilisant les différents classifeurs.

Avec la méthode MFCC-GMM, nous avons pu les classer par rapport aux sujets sains avec un EER de 37%. Mais si on considère seulement les iRBD $moteur^+$, définis par un score moteur UPDRS III > 14, l'EER descend à 28%. Ceci semble indiquer que la méthode d'analyse MFCC-GMM pourrait déceler les pré-parkinsoniens à partir du moment où ils commencent à développer les premiers symptômes moteurs.

Quant à l'analyse des différents paramètres globaux, nous avons pu constater que les moyennes des paramètres liés à la prosodie, et ceux liés aux pauses, étaient à chaque fois comprises entre les moyennes des MP et des sains. Ceci va dans le sens d'une altération progressive et dès le stade prodromique de ces domaines vocaux. Concernant les paramètres rythmiques, seule la variabilité relative du rythme (*RSD*) semble être altérée chez les iRBD. Ceci pourrait signifier que la capacité à garder un rythme constant, et la capacité à répéter un rythme imposé, feraient intervenir des processus physiologiques différents, dont le premier serait altéré plus tôt que le deuxième, dans le développement de la maladie de Parkinson.

La classification des iRBD par rapport aux sujets sains, avec le classifieur SVM (à partir du même ensemble de 6 paramètres) a conduit à un EER de 38% pour les enregistrements du microphone professionnel. Ces résultats sont améliorés si on considère seulement les iRBD $moteur^+$ (EER de 33%).

Corrélations avec la neuroimagerie et la clinique

Les analyses sur les classifications MP débutant vs sain répondaient à un objectif d'aide au diagnostic précoce par l'analyse vocale. Pour répondre à l'objectif complémentaire d'aide au suivi de l'évolution de MP par l'analyse de la voix, nous avons analysé s'il y avait des corrélations

entre des paramètres vocaux et l'état d'avancement de la maladie, quantifié par des résultats de neuroimagerie, et par le score moteur de l'UPDRS III. Pour cela nous avons essayé de prédire les résultats du DatScan, de l'IRM sensible à la neuromélanine et de l'UPDRS III à partir d'un ensemble restreint de 7 paramètres vocaux extraits sur une sous-partie de notre base de données, enregistrée avec le microphone professionnel. A partir de ces paramètres, caractérisant la prosodie, la répartition des pauses et la phonation, nous avons pu expliquer, grâce à un modèle de régression linéaire multiple, 42% de la variance des données du DatScan, 19% de celle des données d'IRM sensible à la neuromélanine et 21% de celle des scores UPDRS III. C'est à dire que ces 7 paramètres, ainsi que l'information du genre, sont capables de prédire linéairement ces données de manière significative. La précision concernant la prédiction des données du DatScan est particulièrement intéressante car ces données sont connues pour caractériser spécialement bien l'évolution de MP, notamment au stade débutant.

Comparaison avec l'état de l'art

Au final, avec notre étude nous arrivons à détecter les hommes MP débutants enregistrés avec le microphone professionnel, avec une précision (Acc) de 89% au seuil EER. Ce taux de bonnes classifications des MP débutants est meilleur que la majorité des articles de la littérature, malgré le fait que nos patients ont été enregistrés sous l'effet de leurs traitements, atténuant les modifications vocales et rendant plus difficile la classification. Concernant les femmes, nous avons obtenu une performance de 70% au seuil EER. Ce taux est difficilement comparable à l'état de l'art, car nous n'avons trouvé aucune étude sur la détection précoce de MP chez les femmes par la voix, la grande majorité des études mélangeant hommes et femmes dans les analyses.

Un autre intérêt de notre travail est que les classifications sont entièrement automatiques, contrairement à plusieurs études de la littérature, requérant des interventions manuelles (comme des segmentations) pour l'analyse.

Enfin la dernière originalité de notre étude, est l'utilisation de la langue française pour la détection de MP. Des travaux antérieurs existaient sur l'analyse de la voix de patients MP en langue française, aucun à notre connaissance n'avait été jusqu'à la classification à partir de cette dernière. La particularité de la langue française est que c'est une langue avec peu d'accents toniques, ce qui atténue les différences entre la parole MP et la parole saine.

A partir des enregistrements téléphoniques, effectués par les participants en autonomie avec leurs propres téléphones, nous sommes parvenus à des performances de classification de 75% pour les hommes et de 67% pour les femmes, aux seuils EER. Ces performances sont légèrement inférieures à celles que nous avons obtenues avec le microphone professionnel, ce qui était attendu vu les qualités réduites des enregistrements. Nous ne pouvons pas comparer ces résultats à l'état de l'art car c'est la première étude portant sur la détection précoce de MP à partir d'enregistrements issus du réseau téléphonique. En effet les études qui se sont intéressées au télédiagnostic de MP par la voix, ont utilisé pour la plupart, des enregistrements effectués à partir de kits audio, ou d'applications smartphones ou tablettes. L'avantage est que la voix ne passe alors pas par le réseau téléphonique et est donc de meilleure qualité. L'inconvénient est que cela nécessite un matériel spécifique pour les sujets (kit audio particulier, ou smartphone avec un système d'exploitation compatible avec l'application), ce qui est plus compliqué pour la généralisation éventuelle de ces méthodes de diagnostic en dehors du cadre de ces études. Enfin quelques études se sont intéressées à l'effet de la transmission de la voix par le réseau téléphonique sur la détection de MP, mais en simulant ce dernier à partir d'enregistrements de haute qualité. Nos analyses sur les enregistrements téléphoniques réels représentent donc un pas important vers un télédiagnostic précoce de la maladie de Parkinson.

Concernant nos analyses de corrélations avec les données de neuroimagerie, nous avons pu

prédire linéairement de manière significative les résultats du DatScan et d'IRM sensible à la NM à partir de paramètres vocaux. Il est cependant difficile de comparer ces résultats à l'état de l'art, car les corrélations entre des paramètres vocaux et l'imagerie caractérisant le changement dans le système dopaminergique (comme le DatScan ou l'IRM sensible à la NM) n'avait à notre connaissance pas encore été étudiées.

Limites et perspectives

Concernant les limites de notre études, plusieurs points sont à mentionner.

Tout d'abord il faut noter que nous nous sommes servis des performances globales de validation pour optimiser certains hyperparamètres (tels que le nombre de gaussiennes et la plage de fréquence pour la méthode MFCC-GMM, la taille des segments pour les x-vecteurs...). De même pour les SVM nous avons choisi quels paramètres globaux garder d'après les analyses de variance faites sur la base de données entière. Ainsi pour les 3 types de classification, nous restons dans le cadre du développement (même si robuste, grâce à la validation croisée) et non dans le cadre de validation finale. Pour connaitre la performance exacte de généralisation, il faudrait tester nos modèles finaux sur une nouvelle base de sujets non vus, et utiliser le seuil choisi lors du développement pour la classification finale.

Une autre limite de notre étude est que les sujets MP de notre base sont traités pharmacologiquement et ont été enregistrés en ON, donc sous l'effet de leur traitement. Or le traitement peut avoir un effet négatif sur la classification, en atténuant certains changements vocaux. Nos modèles de classification seraient à réajuster pour la détection précoce de MP chez des sujets non traités, afin d'avoir les meilleures performances possibles.

D'autre part, les détections de MP que nous avons effectuées ont toujours été faites par comparaison à des sujets sains. Il serait nécessaire de les compléter par des analyses de diagnostic différentiel, visant par exemple à différencier la maladie de Parkinson des autres syndromes parkinsoniens.

Enfin d'autres analyses pourraient être faites pour améliorer les performances de classification, surtout celle des femmes. Tout d'abord d'autres types de paramètres (comme les RASTA-PLP, les énergies Bark-Band, les d-vecteurs, les *Voice Onset Time*, les formants, etc..) pourraient être utilisés en combinaison des paramètres actuels. Ensuite d'autres classifieurs pourraient être testés, notamment ceux qui sont plus spécifiques aux tâches texte-dépendant comme les modèles de Markov cachés, les réseaux de neurones récurrents, etc.. Enfin d'autres types de fusion pourraient avoir lieu, comme par exemple utiliser certaines statistiques concernant les MFCC (moyennes, écarts types, *kurtosis*, *skewness*) en plus des paramètres globaux que nous avons extraits, pour la classification SVM. Pour finir, d'ici peu de temps, la période couverte par les différents enregistrements de nos participants (enregistrés une fois par an à l'hôpital, et une fois par mois par téléphone) sera suffisamment grande pour permettre une étude longitudinale. Suivre l'évolution des paramètres vocaux, ou du score de classification, fournira une information supplémentaire qui devrait permettre d'augmenter les performances de classification.

L'intérêt de poursuivre toutes ces analyses pour améliorer les performances de détection de MP au stade débutant est multiple. Cela permettrait, à terme, aux médecins de compléter leurs examens cliniques par un test vocal rapide s'ils ont un doute sur le commencement d'une maladie de Parkinson. L'avantage d'un tel test serait le faible coût, la rapidité (quelques minutes) et l'objectivité des résultats. Concernant la possible détection par téléphone, cela permettrait d'aiguiller facilement et rapidement les sujets qui se demandent s'ils ne sont pas en train de développer cette maladie. Si le résultat du test est négatif, on peut imaginer qu'un message leur

soit envoyé pour les rassurer et leur dire qu'ils ne présentent pas les caractéristiques vocales de la maladie de Parkinson, et que donc il est peu probable qu'ils soient en train de la développer. Si le test se révèle positif, on peut imaginer un message les informant que d'après l'analyse vocale il y a une suspicion de MP, et qu'il serait pertinent de consulter un neurologue pour confirmer le diagnostic. En effet cela ne serait pas très éthique qu'un outil d'analyse automatique pose un diagnostic formel de MP par téléphone, et de plus, le médecin aura une vue d'ensemble qui peut être complémentaire au test vocal. L'analyse de la voix, en procurant une aide au diagnostic précoce, pourrait permettre d'avancer le diagnostic de quelques années et même d'identifier les personnes à risque dès le stade préclinique. L'intérêt d'avancer le diagnostic est de pouvoir, à terme, commencer un traitement stoppant l'évolution de la maladie le plus tôt possible, quand un tel traitement existera. En attendant, l'intérêt est de pouvoir identifier et inclure des personnes à risque (avec leur consentement) dans des protocoles de recherche médicamenteuse, afin d'évaluer l'efficacité de molécules sur des cerveaux qui ont encore peu de lésions irréversibles.

Même si les acquisitions et les analyses ont déjà été validées par un comité d'éthique et l'ANSM, l'application concrète à des outils de diagnostic devra faire l'objet de futures discussions devant des comités d'éthique.

En ce qui concerne les corrélations des paramètres vocaux avec la neuroimagerie, la prédiction de 42% de la variance des données du DatScan à partir de 7 paramètres vocaux est un bon début, mais non suffisant pour prédire de façon fiable l'état d'avancement de la maladie, et donc suivre son évolution.

L'intérêt de poursuivre l'analyse de corrélations avec la voix serait de pouvoir trouver un ensemble optimal de paramètres vocaux qui permettrait de prédire la quasi-totalité de la quantité de transporteurs dopaminergiques, et donc d'avoir un examen moins couteux et plus accessible pouvant fournir des informations équivalentes, et aider au diagnostic ainsi qu'au suivi de l'évolution de la maladie. Plusieurs améliorations peuvent être apportées à l'ensemble de paramètres vocaux qu'on a utilisé, comme la suppression d'un des deux paramètres liés aux pauses (pour enlever une redondance), et l'ajout d'autres paramètres liés par exemple à l'articulation ou à la capacité de suivre un rythme constant.

Dans cette analyse nous nous sommes intéressés, en ce qui concerne le DatScan, seulement aux corrélations avec la région du cerveau la plus discriminante dans la MP, à savoir la partie bilatérale sensorimotrice du putamen. Or les sous parties limbiques et associatives du putamen ont aussi été analysées par nos collègues de l'hôpital Pitié Salpêtrière, ainsi que d'autres parties du striatum, comme le noyau caudé et le noyau accumbens, segmentés également en sous régions fonctionnelles. Il serait intéressant de tester le pouvoir prédictif de nos paramètres vocaux sur l'ensemble de ces régions, et d'étudier quel type de paramètre corrèle avec quelle région. Comme les paramètres vocaux reflètent chacun certains troubles spécifiques liés à la maladie de Parkinson, cela pourrait nous permettre de mieux comprendre les différentes altérations des circuits neuronaux dans les premiers stades de la maladie de Parkinson.

Également, faire l'analyse de corrélation sur plus de sujets nous permettra de séparer les hommes des femmes, dans les modèles de régression linéaire, afin de mieux comprendre les éventuelles différences de mécanisme d'altération de ces réseaux entre les hommes et les femmes, cf. [Haaxma et al., 2007].

Enfin les corrélations des paramètres vocaux avec d'autres variables comme des scores cognitifs ou des scores génétiques pourraient également s'avérer utiles, afin de pouvoir prévoir, par exemple, certains déclins cognitifs associés à la maladie de Parkinson.

Bibliographie

[Aarsland et al., 2009] Aarsland, D., Brønnick, K., Alves, G., Tysnes, O. B., Pedersen, K. F., Ehrt, U., and Larsen, J. P. (2009). The spectrum of neuropsychiatric symptoms in patients with early untreated Parkinson's disease. *Journal of Neurology, Neurosurgery & Psychiatry*, 80(8) :928–930.

[Abbott et al., 2001] Abbott, R. D., Petrovitch, H., White, L. R., Masaki, K. H., Tanner, C. M., Curb, J. D., Grandinetti, A., Blanchette, P. L., Popper, J. S., and Ross, G. W. (2001). Frequency of bowel movements and the future risk of Parkinson's disease. *Neurology*, 57(3) :456.

[Ahlskog and Muenter, 2001] Ahlskog, J. E. and Muenter, M. D. (2001). Frequency of levodopa-related dyskinesias and motor fluctuations as estimated from the cumulative literature. *Movement Disorders*, 16(3) :448–458.

[Andrews et al., 2000] Andrews, W. D., Campbell, J. P., and Reynolds, D. A. (2000). Bootstrapping for speaker recognition. In *INTERSPEECH*.

[Arnold et al., 2014] Arnold, C., Gehrig, J., Gispert, S., Seifried, C., and Kell, C. A. (2014). Pathomechanisms and compensatory efforts related to Parkinsonian speech. *NeuroImage : Clinical*, 4 :82–97.

[Arnulf, 2012] Arnulf, I. (2012). REM sleep behavior disorder : Motor manifestations and pathophysiology. *Movement Disorders*, 27(6) :677–689.

[Arora et al., 2015] Arora, S., Venkataraman, V., Zhan, A., Donohue, S., Biglan, K. M., Dorsey, E. R., and Little, M. A. (2015). Detecting and monitoring the symptoms of Parkinson's disease using smartphones : A pilot study. *Parkinsonism & related disorders*, 21(6) :650–653.

[Bardinet et al., 2009] Bardinet, E., Bhattacharjee, M., Dormont, D., Pidoux, B., Malandain, G., Schupbach, M., Ayache, N., Cornu, P., Agid, Y., and Yelnik, J. (2009). A three-dimensional histological atlas of the human basal ganglia. II. Atlas deformation strategy and evaluation in deep brain stimulation for Parkinson disease. *Journal of Neurosurgery*, 110(2) :208–19.

[Barone et al., 2009] Barone, P., Antonini, A., Colosimo, C., Marconi, R., Morgante, L., Avarello, T. P., Bottacchi, E., Cannas, A., Ceravolo, G., Ceravolo, R., Cicarelli, G., Gaglio, R. M., Giglia, R. M., Iemolo, F., Manfredi, M., Meco, G., Nicoletti, A., Pederzoli, M., Petrone, A., Pisani, A., Pontieri, F. E., Quatrale, R., Ramat, S., Scala, R., Volpe, G., Zappulla, S., Bentivoglio, A. R., Stocchi, F., Trianni, G., Dotto, P. D., and on behalf of the PRIAMO study group (2009). The PRIAMO study : A multicenter assessment of nonmotor symptoms and their impact on quality of life in Parkinson's disease. *Movement Disorders*, 24(11) :1641–1649.

[Benba et al., 2014] Benba, A., Jilbab, A., and Hammouch, A. (2014). Voice analysis for detecting persons with Parkinson's disease using MFCC and VQ. In *The 2014 international conference on circuits, systems and signal processing*, pages 23–25.

[Benba et al., 2015] Benba, A., Jilbab, A., and Hammouch, A. (2015). Detecting Patients with Parkinson's disease using Mel Frequency Cepstral Coefficients a nd Support Vector Machines. *International Journal on Electrical Engineering and Informatics*, 7(2) :297–307.

[Benba et al., 2016a] Benba, A., Jilbab, A., and Hammouch, A. (2016a). Analysis of multiple types of voice recordings in cepstral domain using MFCC for discriminating between patients

with Parkinson's disease and healthy people. *International Journal of Speech Technology*, 19(3) :449–456.

[Benba et al., 2016b] Benba, A., Jilbab, A., and Hammouch, A. (2016b). Discriminating Between Patients With Parkinson's and Neurological Diseases Using Cepstral Analysis. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 24(10) :1100–1108.

[Benba et al., 2017] Benba, A., Jilbab, A., and Hammouch, A. (2017). Using Human Factor Cepstral Coefficient on Multiple Types of Voice Recordings for Detecting Patients with Parkinson's Disease. *IRBM*, 38(6) :346–351.

[Bimbot et al., 2004] Bimbot, F., Bonastre, J.-F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., Merlin, T., Ortega-García, J., Petrovska-Delacrétaz, D., and Reynolds, D. A. (2004). A Tutorial on Text-Independent Speaker Verification. *EURASIP Journal on Advances in Signal Processing*, 2004(4) :101962.

[Bocklet et al., 2013] Bocklet, T., Steidl, S., Nöth, E., and Skodda, S. (2013). Automatic evaluation of parkinson's speech-acoustic, prosodic and voice related cues. In *Interspeech*, pages 1149–1153.

[Boersma, 1993] Boersma, P. (1993). ACCURATE SHORT-TERM ANALYSIS OF THE FUNDAMENTAL FREQUENCY AND THE HARMONICS-TO-NOISE RATIO OF A SAMPLED SOUND. In *Proceedings of the Institute of Phonetic Sciences*, volume 17, pages 97–110, University of Amsterdam.

[Boersma and Weenink, 2001] Boersma, P. and Weenink, D. (2001). PRAAT, a system for doing phonetics by computer. *Glot international*, 5 :341–345.

[Boll, 1979] Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(2) :113–120.

[Booth et al., 2015] Booth, T. C., Nathan, M., Waldman, A. D., Quigley, A.-M., Schapira, A. H., and Buscombe, J. (2015). The Role of Functional Dopamine-Transporter SPECT Imaging in Parkinsonian Syndromes, Part 1. *American Journal of Neuroradiology*, 36(2) :229–235.

[Braak et al., 2006] Braak, H., de Vos, R. A. I., Bohl, J., and Del Tredici, K. (2006). Gastric alpha-synuclein immunoreactive inclusions in Meissner's and Auerbach's plexuses in cases staged for Parkinson's disease-related brain pathology. *Neuroscience Letters*, 396(1) :67–72.

[Braak et al., 2003] Braak, H., Tredici, K. D., Rüb, U., de Vos, R. A. I., Jansen Steur, E. N. H., and Braak, E. (2003). Staging of brain pathology related to sporadic Parkinson's disease. *Neurobiology of Aging*, 24(2) :197–211.

[Breiman, 1996] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2) :123–140.

[Brunato et al., 2013] Brunato, M., Battiti, R., Pruitt, D., and Sartori, E. (2013). Supervised and unsupervised machine learning for the detection, monitoring and management of Parkinson's disease from passive mobile phone data. In : Predicting Parkinson's Disease Progression with Smartphone Data. *Kaggle Competition. Available at : https ://kaggle2. blob. core. windows. net/prospectorfiles/1117/958625cf-3514-4e64-b0e7-13ebd3cf9791/kaggle. pdf. Accessed October*.

[Brundin et al., 2010] Brundin, P., Melki, R., and Kopito, R. (2010). Prion-like transmission of protein aggregates in neurodegenerative diseases. *Nature Reviews Molecular Cell Biology*, 11(4) :301–307.

[Bühlmann and Yu, 2002] Bühlmann, P. and Yu, B. (2002). Analyzing Bagging. *The Annals of Statistics*, 30(4) :927–961.

[Campbell, 1997] Campbell, J. P. (1997). Speaker recognition : a tutorial. *Proceedings of the IEEE*, 85(9) :1437–1462.

[Campbell et al., 2006] Campbell, W., Sturim, D., and Reynolds, D. (2006). Support vector machines using GMM supervectors for speaker verification. *IEEE Signal Processing Letters*, 13(5) :308–311.

[Chen et al., 2015] Chen, H., Zhao, E. J., Zhang, W., Lu, Y., Liu, R., Huang, X., Ciesielski-Jones, A. J., Justice, M. A., Cousins, D. S., and Peddada, S. (2015). Meta-analyses on prevalence of selected Parkinson's nonmotor symptoms before and after diagnosis. *Translational Neurodegeneration*, 4(1) :1.

[Countryman et al., 2003] Countryman, S., Camburn, J., and Schwantz, J. (2003). *Parkinson's Disease : Speaking Out*. National Parkinson Foundation, 6 edition.

[Darley Frederic L. et al., 1969] Darley Frederic L., Aronson Arnold E., and Brown Joe R. (1969). Differential Diagnostic Patterns of Dysarthria. *Journal of Speech and Hearing Research*, 12(2) :246–269.

[Davis and Mermelstein, 1980] Davis, S. and Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 28(4) :357–366.

[De Cock et al., 2007] De Cock, V. C., Vidailhet, M., Leu, S., Texeira, A., Apartis, E., Elbaz, A., Roze, E., Willer, J. C., Derenne, J. P., Agid, Y., and Arnulf, I. (2007). Restoration of normal motor control in Parkinson's disease during REM sleep. *Brain*, 130(2) :450–456.

[De Lau and Breteler, 2006] De Lau, L. M. and Breteler, M. M. (2006). Epidemiology of Parkinson's disease. *The Lancet Neurology*, 5(6) :525–535.

[de Lima Xavier et al., 2019] de Lima Xavier, L., Hanekamp, S., and Simonyan, K. (2019). Sexual Dimorphism Within Brain Regions Controlling Speech Production. *Frontiers in Neuroscience*, 13.

[Dehak et al., 2011] Dehak, N., Kenny, P., Dehak, R., Dumouchel, P., and Ouellet, P. (2011). Front End Factor Analysis For Speaker Verification. *IEEE TRANSACTIONS ON AUDIO*, page 13.

[Dibazar et al., 2002] Dibazar, A. A., Narayanan, S., and Berger, T. W. (2002). Feature analysis for automatic detection of pathological speech. In *Proceedings of the Second Joint 24th Annual Conference and the Annual Fall Meeting of the Biomedical Engineering Society] [Engineering in Medicine and Biology*, volume 1, pages 182–183 vol.1.

[Dorsey et al., 2007] Dorsey, E. R., Constantinescu, R., Thompson, J. P., Biglan, K. M., Holloway, R. G., Kieburtz, K., Marshall, F. J., Ravina, B. M., Schifitto, G., Siderowf, A., and Tanner, C. M. (2007). Projected number of people with Parkinson disease in the most populous nations, 2005 through 2030. *Neurology*, 68(5) :384–386.

[Drissi et al., 2019] Drissi, T. B., Zayrit, S., Nsiri, B., and Ammoummou, A. (2019). Diagnosis of Parkinson's Disease based on Wavelet Transform and Mel Frequency Cepstral Coefficients. *International Journal of Advanced Computer Science and Applications*, 10(3).

[Elgh et al., 2009] Elgh, E., Domellöf, M., Linder, J., Edström, M., Stenlund, H., and Forsgren, L. (2009). Cognitive function in early Parkinson's disease : a population-based study. *European Journal of Neurology*, 16(12) :1278–1284.

[Fearnley and Lees, 1991] Fearnley, J. M. and Lees, A. J. (1991). Ageing and Parkinson's disease : substantia nigra regional selectivity. *Brain : A Journal of Neurology*, 114 (Pt 5) :2283–2301.

[Fraile et al., 2009a] Fraile, R., Godino-Llorente, J. I., Saenz-Lechon, N., Osma-Ruiz, V., and Fredouille, C. (2009a). MFCC-based remote pathology detection on speech transmitted through the telephone channel. *Proc Biosignals*.

[Fraile et al., 2009b] Fraile, R., Sáenz-Lechón, N., Godino-Llorente, J., Osma-Ruiz, V., and Fredouille, C. (2009b). Automatic Detection of Laryngeal Pathologies in Records of Sustained Vowels by Means of Mel-Frequency Cepstral Coefficient Parameters and Differentiation of Patients by Sex. *Folia Phoniatrica et Logopaedica*, 61(3) :146–152.

[Friedman et al., 2001] Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

[Gabriel, 2019] Gabriel, C. (2019). Production de la parole et voix humaine.

[Garcia-Ospina et al., 2018] Garcia-Ospina, N., Arias-Vergara, T., Vásquez-Correa, J. C., Orozco-Arroyave, J. R., Cernak, M., and Nöth, E. (2018). Phonological i-Vectors to Detect Parkinson's Disease. In Sojka, P., Horák, A., Kopeček, I., and Pala, K., editors, *Text, Speech, and Dialogue*, Lecture Notes in Computer Science, pages 462–470. Springer International Publishing.

[Gaurav et al., 2019] Gaurav, R., Yahia-Cherif, L., Mangone, G., Pyatigorskaya, N., Valabregue, R., Ewenczyk, C., Hutchison, M., Vidailhet, M., and Lehéricy, S. (2019). LONGITUDINAL VARIATIONS IN NEUROMELANIN MRI SIGNAL IN PARKINSON'S DISEASE.

[Ghio and Pinto, 2007] Ghio, A. and Pinto, S. (2007). Résonance sonore et cavités supralaryngée. In *Les Dysarthries*, pages 101–110. SOLAL.

[Gil and Johnson, 2009] Gil, D. and Johnson, M. (2009). Diagnosing Parkinson by using Artificial Neural Networks and Support Vector Machines. *Global Journal of Computer Science and Technology*, 9.

[Godino-Llorente et al., 2006] Godino-Llorente, J., Gomez-Vilda, P., and Blanco-Velasco, M. (2006). Dimensionality Reduction of a Pathological Voice Quality Assessment System Based on Gaussian Mixture Models and Short-Term Cepstral Parameters. *IEEE Transactions on Biomedical Engineering*, 53(10) :1943–1953.

[Godino-Llorente and Gómez-Vilda, 2004] Godino-Llorente, J. and Gómez-Vilda, P. (2004). Automatic Detection of Voice Impairments by Means of Short-Term Cepstral Parameters and Neural Network Based Detectors. *IEEE Transactions on Biomedical Engineering*, 51(2) :380–384.

[Goetz et al., 2007] Goetz, C. G., Fahn, S., Martinez-Martin, P., Poewe, W., Sampaio, C., Stebbins, G. T., Stern, M. B., Tilley, B. C., Dodel, R., Dubois, B., Holloway, R., Jankovic, J., Kulisevsky, J., Lang, A. E., Lees, A., Leurgans, S., LeWitt, P. A., Nyenhuis, D., Olanow, C. W., Rascol, O., Schrag, A., Teresi, J. A., Hilten, J. J. V., and LaPelle, N. (2007). Movement Disorder Society-sponsored revision of the Unified Parkinson's Disease Rating Scale (MDS-UPDRS) : Process, format, and clinimetric testing plan. *Movement Disorders*, 22(1) :41–47.

[Goldman and Holden, 2016] Goldman, J. and Holden, S. (2016). Parkinson's Disease. In *Encyclopedia of Mental Health*, pages 242–248. Elsevier.

[Graff et al., 1998] Graff, D., Canavan, A., and Zipperlen, G. (1998). Switchboard-2 Phase I.

[Graff et al., 1999] Graff, D., Walker, K., and Canavan, A. (1999). Switchboard-2 Phase II.

[Graff et al., 2001] Graff, D., Walker, K., and Miller, D. (2001). Switchboard Cellular Part 1 Audio.

[Graff et al., 2004] Graff, D., Walker, K., and Miller, D. (2004). Switchboard Cellular Part 2 Audio.

[Grosz et al., 2015] Grosz, T., Busa-Fekete, R., Gosztolya, G., and Toth, L. (2015). Assessing the Degree of Nativeness and Parkinson's Condition Using Gaussian Processes and Deep Rectifier Neural Networks. In *Interspeech 2015*, page 5.

[Gómez-Vilda et al., 2017] Gómez-Vilda, P., Mekyska, J., Ferrández, J. M., Palacios-Alonso, D., Gómez-Rodellar, A., Rodellar-Biarge, V., Galaz, Z., Smekal, Z., Eliasova, I., Kostalova, M., and Rektorova, I. (2017). Parkinson Disease Detection from Speech Articulation Neuromechanics. *Frontiers in Neuroinformatics*, 11.

[Haas et al., 2012] Haas, B. R., Stewart, T. H., and Zhang, J. (2012). Premotor biomarkers for Parkinson's disease-a promising direction of research. *Transl Neurodegener*, 1(1) :11.

[Haaxma et al., 2007] Haaxma, C. A., Bloem, B. R., Borm, G. F., Oyen, W. J. G., Leenders, K. L., Eshuis, S., Booij, J., Dluzen, D. E., and Horstink, M. W. I. M. (2007). Gender differences in Parkinson's disease. *Journal of Neurology, Neurosurgery &amp ; Psychiatry*, 78(8) :819–824.

[Halawani and Ahmad, 2012] Halawani, S. M. and Ahmad, A. (2012). Ensemble methods for prediction of Parkinson disease. In *Intelligent Data Engineering and Automated Learning-IDEAL 2012*, pages 516–521. Springer.

[Harel et al., 2004] Harel, B., Cannizzaro, M., and Snyder, P. J. (2004). Variability in fundamental frequency during speech in prodromal and incipient Parkinson's disease : A longitudinal case study. *Brain and Cognition*, 56(1) :24–29.

[Hawkes et al., 1997] Hawkes, C. H., Shephard, B. C., and Daniel, S. E. (1997). Olfactory dysfunction in Parkinson's disease. *Journal of Neurology, Neurosurgery, and Psychiatry*, 62(5) :436–446.

[Heigold et al., 2016] Heigold, G., Moreno, I., Bengio, S., and Shazeer, N. (2016). End-to-end text-dependent speaker verification. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5115–5119.

[Hemmerling et al., 2016] Hemmerling, D., Orozco-Arroyave, J. R., Skalski, A., Gajda, J., and Nöth, E. (2016). Automatic Detection of Parkinson's Disease Based on Modulated Vowels. In *INTERSPEECH*, pages 1190–1194.

[Hoehn and Yahr, 1967] Hoehn, M. and Yahr, M. D. (1967). Parkinsonism : onset, progression and mortality. *Neurology*, 17(5) :427–442.

[Hornykiewicz, 1998] Hornykiewicz, O. (1998). Biochemical aspects of Parkinson's disease. *Neurology*, 51(2 Suppl 2) :S2.

[Hughes et al., 1992] Hughes, A. J., Daniel, S. E., Kilford, L., and Lees, A. J. (1992). Accuracy of clinical diagnosis of idiopathic Parkinson's disease : a clinico-pathological study of 100 cases. *Journal of Neurology, Neurosurgery & Psychiatry*, 55(3) :181–184.

[Huh et al., 2015] Huh, Y. E., Park, J., Suh, M. K., Lee, S. E., Kim, J., Jeong, Y., Kim, H.-T., and Cho, J. W. (2015). Differences in early speech patterns between Parkinson variant of multiple system atrophy and Parkinson's disease. *Brain and Language*, 147 :14–20.

[Iansek et al., 1995] Iansek, R., Bradshaw, J. L., Phillips, J. G., Cunnington, R., and Morris, M. E. (1995). Interaction of the basal ganglia and supplementary motor area in the elaboration of movement. In Piek, D. J. G. a. J. P., editor, *Advances in Psychology*, volume 111 of *Motor Control and Sensory Motor IntegrationIssues and Directions*, pages 37–59. North-Holland.

[Iranzo et al., 2014] Iranzo, A., Fernández-Arcos, A., Tolosa, E., Serradell, M., Molinuevo, J. L., Valldeoriola, F., Gelpi, E., Vilaseca, I., Sánchez-Valle, R., Lladó, A., Gaig, C., and Santamaría, J. (2014). Neurodegenerative Disorder Risk in Idiopathic REM Sleep Behavior Disorder : Study in 174 Patients. *PLoS ONE*, 9(2).

[Jafari, 2013] Jafari, A. (2013). Classification of Parkinson's Disease Patients using Nonlinear Phonetic Features and Mel-Frequency Cepstral Analysis. *Biomedical Engineering : Applications, Basis and Communications*, 25(04) :1350001.

[Jankovic, 2003] Jankovic, J. (2003). Pathophysiology And Clinical Assessment Of Parkinsonian Symptoms And Signs. In Lyons, K., Pahwa, R., and Roller, W., editors, *Handbook of Parkinson's Disease, Fourth Edition*, volume 20035941. Informa Healthcare.

[Jeancolas et al., 2017] Jeancolas, L., Benali, H., Benkelfat, B. E., Mangone, G., Corvol, J. C., Vidailhet, M., Lehericy, S., and Petrovska-Delacrétaz, D. (2017). Automatic detection of early stages of Parkinson's disease through acoustic voice analysis with mel-frequency cepstral coefficients. In *2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, pages 1–6.

[Jeancolas et al., 2019a] Jeancolas, L., Mangone, G., Corvol, J.-C., Vidailhet, M., Lehéricy, S., Benkelfat, B.-E., Benali, H., and Petrovska-Delacrétaz, D. (2019a). Comparison of Telephone Recordings and Professional Microphone Recordings for Early Detection of Parkinson's Disease, Using Mel-Frequency Cepstral Coefficients with Gaussian Mixture Models. In *Interspeech 2019*, pages 3033–3037. ISCA.

[Jeancolas et al., 2019b] Jeancolas, L., Mangone, G., Villain, N., Gaurav, R., Habert, M. O., Corvol, J. C., Vidailhet, M., Benkelfat, B.-E., Lehéricy, S., Benali, H., and Petrovska-Delacrétaz, D. (2019b). Analyse de la Voix au Stade Débutant de la Maladie de Parkinson et Corrélations avec Analyse clinique et Neuroimagerie. In *Journées d'Etude sur la TéléSanté*, Paris, France. Sorbonne Universités.

[Jeancolas et al., 2016] Jeancolas, L., Petrovska-Delacrétaz, D., Lehéricy, S., Benali, H., and Benkelfat, B.-E. (2016). L'analyse de la voix comme outil de diagnostic précoce de la maladie de Parkinson : état de l'art. In *CORESA 2016 : 18e Edition COmpressions et REprésentation des Signaux Audiovisuels*, pages 113–121, Nancy. CNRS.

[Jiang et al., 1999] Jiang, J., Lin, E., Wang, J., and Hanson, D. G. (1999). Glottographic Measures Before and After Levodopa Treatment in Parkinson's Disease. *The Laryngoscope*, 109(8) :1287–1294.

[Jung et al., 2019] Jung, M., Mody, M., Fujioka, T., Kimura, Y., Okazawa, H., and Kosaka, H. (2019). Sex Differences in White Matter Pathways Related to Language Ability. *Frontiers in Neuroscience*, 13.

[Kapoor and Sharma, 2011] Kapoor, T. and Sharma, R. K. (2011). Parkinson's disease diagnosis using Mel-frequency cepstral coefficients and vector quantization. *International Journal of Computer Applications*, 14(3) :43–46.

[Kenny et al., 2005] Kenny, P., Boulianne, G., and Dumouchel, P. (2005). Eigenvoice modeling with sparse training data. *IEEE Transactions on Speech and Audio Processing*, 13(3) :345–354.

[Kenny et al., 2007] Kenny, P., Boulianne, G., Ouellet, P., and Dumouchel, P. (2007). Joint Factor Analysis Versus Eigenchannels in Speaker Recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 15(4) :1435–1447.

[Kenny et al., 2003] Kenny, P., Mihoubi, M., and Dumouchel, P. (2003). New MAP Estimators for Speaker Recognition. In *INTERSPEECH*, page 4.

[Kharroubi et al., 2001] Kharroubi, J., Petrovska-Delacretaz, D., and Chollet, G. (2001). Combining GMM's with Suport Vector Machines for Text-independent Speaker Verification. In *Eurospeech 2001*, page 5, Scandinavia.

[Khojasteh et al., 2018] Khojasteh, P., Viswanathan, R., Aliahmad, B., Ragnav, S., Zham, P., and Kumar, D. K. (2018). Parkinson's Disease Diagnosis Based on Multivariate Deep Features of Speech Signal. In *2018 IEEE Life Sciences Conference (LSC)*, pages 187–190.

[Kibleur, 2016] Kibleur, A. (2016). *Cartographie corticale par électroencéphalographie des effets de la stimulation cérébrale profonde chez les patients souffrant de troubles psychiatriques réfractaires et les patients parkinsoniens.* PhD thesis.

[Koenig, 2011] Koenig, L. (2011). *Masquage de pertes de paquets en voix sur IP*. PhD thesis.

[Kohavi, 1995] Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145. Stanford, CA.

[Kulisevsky et al., 2008] Kulisevsky, J., Pagonabarraga, J., Pascual-Sedano, B., García-Sánchez, C., Gironell, A., and Trapecio Group Study (2008). Prevalence and correlates of neuropsychiatric symptoms in Parkinson's disease without dementia. *Movement Disorders*, 23(13) :1889–1896.

[Kyung and Lee, 1999] Kyung, Y. J. and Lee, H. S. (1999). Bootstrap and aggregating VQ classifier for speaker recognition. *Electronics Letters*, 35(12) :973–974.

[Larcher, 2018] Larcher, A. (2018). *Modèles acoustiques pour la reconnaissance du locuteur*. Habilitation à diriger des recherches, Université du Mans.

[Lebouvier et al., 2010] Lebouvier, T., Neunlist, M., Bruley des Varannes, S., Coron, E., Drouard, A., N'Guyen, J.-M., Chaumette, T., Tasselli, M., Paillusson, S., Flamand, M., Galmiche, J.-P., Damier, P., and Derkinderen, P. (2010). Colonic Biopsies to Assess the Neuropathology of Parkinson's Disease and Its Relationship with Symptoms. *PLoS ONE*, 5(9).

[Lei et al., 2014] Lei, Y., Scheffer, N., Ferrer, L., and McLaren, M. (2014). A NOVEL SCHEME FOR SPEAKER RECOGNITION USING A PHONETICALLY-AWARE DEEP NEURAL NETWORK. In *IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP*, page 5.

[Li and Zheng, 2015] Li, L. and Zheng, T. F. (2015). Gender-dependent feature extraction for speaker recognition. In *2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP)*, pages 509–513.

[Li et al., 2010] Li, S. X., Wing, Y. K., Lam, S. P., Zhang, J., Yu, M. W. M., Ho, C. K. W., Tsoh, J., and Mok, V. (2010). Validation of a new REM sleep behavior disorder questionnaire (RBDQ-HK). *Sleep Medicine*, 11(1) :43–48.

[Lim et al., 2009] Lim, S.-Y., Fox, S. H., and Lang, A. E. (2009). Overview of the Extranigral Aspects of Parkinson Disease. *ARCH NEUROL*, 66(2) :6.

[Little et al., 2009] Little, M., McSharry, P., Hunter, E., Spielman, J., and Ramig, L. (2009). Suitability of Dysphonia Measurements for Telemonitoring of Parkinson's Disease. *IEEE Transactions on Biomedical Engineering*, 56(4) :1015–1022.

[Locco, 2005] Locco, J. (2005). *La production des occlusives dans la maladie de Parkinson*. PhD thesis, Aix-Marseilles.

[Maillard et al., 2017] Maillard, G., Arlot, S., and Lerasle, M. (2017). Cross-validation improved by aggregation : Agghoo. *hal*, page 21.

[Malyska et al., 2005] Malyska, N., Quatieri, T. F., and Sturim, D. (2005). Automatic dysphonia recognition using biologically-inspired amplitude-modulation features. In *Acoustics, Speech, and Signal Processing, 2005. Proceedings.(ICASSP'05). IEEE International Conference on*, volume 1, pages I–873. IEEE.

[Meissner, 2012] Meissner, W. (2012). When does Parkinson's disease begin ? From prodromal disease to motor signs. *Revue Neurologique*, 168(11) :809–814.

[Meunier, 2007] Meunier, C. (2007). Phonétique acoustique. In *Les Dysarthries*, pages 164–173. SOLAL.

[Moisan, 2018] Moisan, F. (2018). FRÉQUENCE DE LA MALADIE DE PARKINSON EN FRANCE EN 2015 ET ÉVOLUTION JUSQU'EN 2030. *Bulletin Epidémiologique Hebdomadaire*, 8-9 :13.

[Moro-Velázquez et al., 2018] Moro-Velázquez, L., Gómez-García, J. A., Godino-Llorente, J. I., Villalba, J., Orozco-Arroyave, J. R., and Dehak, N. (2018). Analysis of speaker recognition methodologies and the influence of kinetic changes to automatically detect Parkinson's Disease. *Applied Soft Computing*, 62 :649–666.

[Murphy, 2002] Murphy, C. (2002). Prevalence of Olfactory Impairment in Older Adults. *JAMA*, 288(18) :2307.

[Müller J et al., 2001] Müller J, Wenning GK, Verny M, and et al (2001). Progression of dysarthria and dysphagia in postmortem-confirmed parkinsonian disorders. *Archives of Neurology*, 58(2) :259–264.

[Nagrani et al., 2017] Nagrani, A., Chung, J. S., and Zisserman, A. (2017). VoxCeleb : A Large-Scale Speaker Identification Dataset. In *Interspeech 2017*, pages 2616–2620. ISCA.

[Narayana et al., 2010] Narayana, S., Fox, P. T., Zhang, W., Franklin, C., Robin, D. A., Vogel, D., and Ramig, L. O. (2010). Neural correlates of efficacy of voice therapy in Parkinson's disease identified by performance–correlation analysis. *Human Brain Mapping*, 31(2) :222–236.

[Narayana et al., 2009] Narayana, S., Jacks, A., Robin, D. A., Poizner, H., Zhang, W., Franklin, C., Liotti, M., Vogel, D., and Fox, P. T. (2009). A Non-Invasive Imaging Approach to Understanding Speech Changes following Deep Brain Stimulation in Parkinson's Disease. *American journal of speech-language pathology / American Speech-Language-Hearing Association*, 18(2) :146–161.

[Novotný et al., 2014] Novotný, M., Rusz, J., Cmejla, R., and Ruzicka, E. (2014). Automatic Evaluation of Articulatory Disorders in Parkinson's Disease. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(9) :1366–1378.

[Orozco-Arroyave et al., 2014a] Orozco-Arroyave, J. R., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Gonzalez-Rátiva, M. C., and Nöth, E. (2014a). New Spanish Speech Corpus Database for the Analysis of People Suffering from Parkinson's Disease. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

[Orozco-Arroyave et al., 2015a] Orozco-Arroyave, J. R., Belalcazar-Bolaños, E. A., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Skodda, S., Rusz, J., Daqrouq, K., Hönig, F., and Nöth, E. (2015a). Characterization Methods for the Detection of Multiple Voice Disorders : Neurological, Functional, and Laryngeal Diseases. *IEEE Journal of Biomedical and Health Informatics*, 19(6) :1820–1828.

[Orozco-Arroyave et al., 2014b] Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., Bonilla, J. F. V., Skodda, S., Rusz, J., and Nöth, E. (2014b). Automatic detection of Parkinson's disease from words uttered in three different languages. In *INTERSPEECH*, pages 1573–1577.

[Orozco-Arroyave et al., 2015b] Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., Bonilla, J. F. V., Skodda, S., Rusz, J., and Nöth, E. (2015b). Voiced/unvoiced transitions in speech as a potential bio-marker to detect parkinson's disease. In *INTERSPEECH*, pages 95–99. Citeseer.

[Orozco-Arroyave et al., 2016a] Orozco-Arroyave, J. R., Hönig, F., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Daqrouq, K., Skodda, S., Rusz, J., and Nöth, E. (2016a). Automatic detection of Parkinson's disease in running speech spoken in three different languages. *The Journal of the Acoustical Society of America*, 139(1) :481–500.

[Orozco-Arroyave et al., 2016b] Orozco-Arroyave, J. R., Vàsquez-Correa, J. C., Honig, F., Arias-Londono, J. D., Vargas-Bonilla, J. F., Skodda, S., Rusz, J., and Noth, E. (2016b). Towards an automatic monitoring of the neurological state of Parkinson's patients from speech. pages 6490–6494. IEEE.

[Ozkan, 2016] Ozkan, H. (2016). A Comparison of Classification Methods for Telediagnosis of Parkinson's Disease. *Entropy*, 18(4) :115.

[Parveen and Qadeer, 2000] Parveen, S. and Qadeer, A. (2000). SPEAKER RECOGNITION WITH RECURRENT NEURAL NETWORKS. In *ICSLP*, page 4, China.

[Pedersen et al., 2009] Pedersen, K. F., Larsen, J. P., Alves, G., and Aarsland, D. (2009). Prevalence and clinical correlates of apathy in Parkinson's disease : A community-based study. *Parkinsonism & Related Disorders*, 15(4) :295–299.

[Peelaerts et al., 2015] Peelaerts, W., Bousset, L., Van der Perren, A., Moskalyuk, A., Pulizzi, R., Giugliano, M., Van Den Haute, C., Melki, R., and Baekelandt, V. (2015). Alpha-Synuclein

strains cause distinct synucleinopathies after local and systemic administration. *Nature*, 522(7556) :340–344.

[Pelecanos and Sridharan, 2001] Pelecanos, J. W. and Sridharan, S. (2001). Feature warping for robust speaker verification. In *Odyssey*.

[Peppas et al., 2008] Peppas, G., Alexiou, V. G., Mourtzoukou, E., and Falagas, M. E. (2008). Epidemiology of constipation in Europe and Oceania : a systematic review. *BMC Gastroenterology*, 8 :5.

[Pieri et al., 2000] Pieri, V., Diederich, N. J., Raman, R., and Goetz, C. G. (2000). Decreased color discrimination and contrast sensitivity in Parkinson's disease. *Journal of the Neurological Sciences*, 172(1) :7–11.

[Pinto et al., 2014] Pinto, S., Ferraye, M., Espesser, R., Fraix, V., Maillet, A., Guirchoum, J., Layani-Zemour, D., Ghio, A., Chabardès, S., Pollak, P., and Debû, B. (2014). Stimulation of the pedunculopontine nucleus area in Parkinson's disease : effects on speech and intelligibility. *Brain*, 137(10) :2759–2772.

[Pinto et al., 2010] Pinto, S., Ghio, A., Teston, B., and Viallet, F. (2010). La dysarthrie au cours de la maladie de Parkinson. Histoire naturelle de ses composantes : dysphonie, dysprosodie et dysarthrie. *Revue Neurologique*, 166(10) :800–810.

[Ponsen et al., 2004] Ponsen, M. M., Stoffers, D., Booij, J., van Eck-Smit, B. L. F., Wolters, E. C., and Berendse, H. W. (2004). Idiopathic hyposmia as a preclinical sign of Parkinson's disease. *Annals of Neurology*, 56(2) :173–181.

[Postuma, 2015] Postuma, R. B. (2015). Voice changes in prodromal Parkinson's disease - is a new biomarker within earshot ? *Sleep Medicine*.

[Postuma et al., 2012] Postuma, R. B., Lang, A. E., Gagnon, J. F., Pelletier, A., and Montplaisir, J. Y. (2012). How does parkinsonism start ? Prodromal parkinsonism motor changes in idiopathic REM sleep behaviour disorder. *Brain*, 135(6) :1860–1870.

[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., and Vesely, K. (2011). The Kaldi Speech Recognition Toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, page 4.

[Prescrire, 2015] Prescrire (2015). Signes de maladie de Parkinson idiopathique. *La Revue Prescrire*, 35(381) :526–529.

[Prince, 2007] Prince, S. J. D. (2007). Probabilistic Linear Discriminant Analysis for. In *Inferences About Identity ," ICCV*.

[Pyatigorskaya et al., 2018] Pyatigorskaya, N., Magnin, B., Mongin, M., Yahia-Cherif, L., Valabregue, R., Arnaldi, D., Ewenczyk, C., Poupon, C., Vidailhet, M., and Lehéricy, S. (2018). Comparative Study of MRI Biomarkers in the Substantia Nigra to Discriminate Idiopathic Parkinson Disease. *American Journal of Neuroradiology*.

[Ramig et al., 2001] Ramig, L. O., Sapir, S., Countryman, S., Pawlas, A. A., O'brien, C., Hoehn, M., and Thompson, L. L. (2001). Intensive voice treatment (LSVT®) for patients with Parkinson's disease : a 2 year follow up. *Journal of Neurology, Neurosurgery & Psychiatry*, 71(4) :493–498.

[Rektorova et al., 2012] Rektorova, I., Mikl, M., Barrett, J., Marecek, R., Rektor, I., and Paus, T. (2012). Functional neuroanatomy of vocalization in patients with Parkinson's disease. *Journal of the Neurological Sciences*, 313(1–2) :7 – 12.

[Reynolds, 1992] Reynolds, D. A. (1992). A Gaussian mixture modeling approach to text-independent speaker identification.

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

[Reynolds and Rose, 1995] Reynolds, D. A. and Rose, R. C. (1995). Robust text-independent speaker identification using Gaussian mixture speaker models. *IEEE Trans. Speech and Audio Processing*, 3 :72–83.

[Ribeiro et al., 2002] Ribeiro, M.-J., Vidailhet, M., Loc'h, C., Dupel, C., Nguyen, J. P., Ponchant, M., Dollé, F., Peschanski, M., Hantraye, P., Cesaro, P., Samson, Y., and Remy, P. (2002). Dopaminergic Function and Dopamine Transporter Binding Assessed With Positron Emission Tomography in Parkinson Disease. *Archives of Neurology*, 59(4) :580–586.

[Richard, 2016] Richard, G. (2016). Analyse des signaux audiofréquences - Indexation audio.

[Rodriguez-Oroz et al., 2009] Rodriguez-Oroz, M. C., Jahanshahi, M., Krack, P., Litvan, I., Macias, R., Bezard, E., and Obeso, J. A. (2009). Initial clinical manifestations of Parkinson's disease : features and pathophysiological mechanisms. *The Lancet Neurology*, 8(12) :1128–1139.

[Rusz et al., 2011a] Rusz, J., Cmejla, R., Ruzickova, H., and Ruzicka, E. (2011a). Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated Parkinson's disease. *The Journal of the Acoustical Society of America*, 129(1) :350.

[Rusz et al., 2013a] Rusz, J., Cmejla, R., Tykalova, T., Ruzickova, H., Klempir, J., Majerova, V., Picmausova, J., Roth, J., and Ruzicka, E. (2013a). Imprecise vowel articulation as a potential early marker of Parkinson's disease : Effect of speaking task. *The Journal of the Acoustical Society of America*, 134(3) :2171–2181.

[Rusz et al., 2015] Rusz, J., Hlavnička, J., Tykalová, T., Bušková, J., Ulmanová, O., Růžička, E., and Šonka, K. (2015). Quantitative assessment of motor speech abnormalities in idiopathic rapid eye movement sleep behaviour disorder. *Sleep Medicine*.

[Rusz et al., 2018] Rusz, J., Hlavnička, J., Tykalová, T., Novotný, M., Dušek, P., Šonka, K., and Růžička, E. (2018). Smartphone Allows Capture of Speech Abnormalities Associated With High Risk of Developing Parkinson's Disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(8) :1495–1507.

[Rusz et al., 2011b] Rusz, J., Čmejla, R., Růžičková, H., Klempíř, J., Majerová, V., Picmausová, J., Roth, J., and Růžička, E. (2011b). Acoustic assessment of voice and speech disorders in Parkinson's disease through quick vocal test. *Movement Disorders*, 26(10) :1951–1952.

[Rusz et al., 2013b] Rusz, J., Čmejla, R., Růžičková, H., Klempíř, J., Majerová, V., Picmausová, J., Roth, J., and Růžička, E. (2013b). Evaluation of speech impairment in early stages of Parkinson's disease : a prospective study with the role of pharmacotherapy. *Journal of Neural Transmission*, 120(2) :319–329.

[Sachin et al., 2008] Sachin, S., Senthil Kumaran, S., Singh, S., Goyal, V., Shukla, G., Mahajan, H., and Behari, M. (2008). Functional mapping in PD and PSP for sustained phonation and phoneme tasks. *Journal of the Neurological Sciences*, 273(1-2) :51–56.

[Sadjadi et al., 2017] Sadjadi, S. O., Kheyrkhah, T., Tong, A., Greenberg, C. S., Reynolds, D., Singer, E., Mason, L., and Hernandez-Cordero, J. (2017). The 2016 NIST Speaker Recognition Evaluation. In *INTERSPEECH*.

[Sakar et al., 2013] Sakar, B., Isenkul, M., Sakar, C., Sertbas, A., Gurgen, F., Delil, S., Apaydin, H., and Kursun, O. (2013). Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings. *IEEE Journal of Biomedical and Health Informatics*, 17(4) :828–834.

[Sakar et al., 2017] Sakar, B. E., Serbes, G., and Sakar, C. O. (2017). Analyzing the effectiveness of vocal features in early telediagnosis of Parkinson's disease. *PLOS ONE*, 12(8) :e0182428.

[Santamaria et al., 1986] Santamaria, J., Tolosa, E., and Valles, A. (1986). Parkinson's disease with depression. *Neurology*, 36(8) :1130.

[Schenck et al., 2013] Schenck, C. H., Boeve, B. F., and Mahowald, M. W. (2013). Delayed emergence of a parkinsonian disorder or dementia in 81% of older males initially diagnosed with idiopathic REM sleep behavior disorder (RBD) : 16 year update on a previously reported series.

[Schuller et al., 2015] Schuller, B., Steidl, S., Batliner, A., Hantke, S., Honig, F., Orozco-Arroyave, J. R., Noth, E., Zhang, Y., and Weninger, F. (2015). The INTERSPEECH 2015 Computational Paralinguistics Challenge : Nativeness, Parkinson's & Eating Condition. In *INTERSPEECH*, page 5.

[Skodda, 2015] Skodda, S. (2015). Steadiness of syllable repetition in early motor stages of Parkinson's disease. *Biomedical Signal Processing and Control*, 17 :55–59.

[Skodda et al., 2012] Skodda, S., Grönheit, W., and Schlegel, U. (2012). Impairment of Vowel Articulation as a Possible Marker of Disease Progression in Parkinson's Disease. *PLoS ONE*, 7(2) :e32132.

[Skodda et al., 2013] Skodda, S., Lorenz, J., and Schlegel, U. (2013). Instability of syllable repetition in Parkinson's disease—Impairment of automated speech performance ? *Basal Ganglia*, 3(1) :33–37.

[Snyder et al., 2018a] Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., and Khudanpur, S. (2018a). Spoken Language Recognition using X-vectors. In *Odyssey 2018 The Speaker and Language Recognition Workshop*, pages 105–111. ISCA.

[Snyder et al., 2017] Snyder, D., Garcia-Romero, D., Povey, D., and Khudanpur, S. (2017). Deep Neural Network Embeddings for Text-Independent Speaker Verification. In *Interspeech 2017*, pages 999–1003. ISCA.

[Snyder et al., 2018b] Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., and Khudanpur, S. (2018b). X-Vectors : Robust DNN Embeddings for Speaker Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333, Calgary, AB. IEEE.

[Snyder et al., 2016] Snyder, D., Ghahremani, P., Povey, D., Garcia-Romero, D., Carmiel, Y., and Khudanpur, S. (2016). Deep neural network-based speaker embeddings for end-to-end speaker verification. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 165–170, San Diego, CA. IEEE.

[Soong et al., 1985] Soong, F., Rosenberg, A., Rabiner, L., and Juang, B. (1985). A vector quantization approach to speaker recognition. In *ICASSP '85. IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 10, pages 387–390, Tampa, FL, USA. Institute of Electrical and Electronics Engineers.

[Sulzer et al., 2018] Sulzer, D., Cassidy, C., Horga, G., Kang, U. J., Fahn, S., Casella, L., Pezzoli, G., Langley, J., Hu, X. P., Zucca, F. A., Isaias, I. U., and Zecca, L. (2018). Neuromelanin detection by magnetic resonance imaging (MRI) and its promise as a biomarker for Parkinson's disease. *npj Parkinson's Disease*, 4(1).

[Sáenz-Lechón et al., 2006] Sáenz-Lechón, N., Godino-Llorente, J. I., Osma-Ruiz, V., and Gómez-Vilda, P. (2006). Methodological issues in the development of automatic systems for voice pathology detection. *Biomedical Signal Processing and Control*, 1(2) :120–128.

[Tissingh et al., 2001] Tissingh, G., Berendse, H. W., Bergmans, P., DeWaard, R., Drukarch, B., Stoof, J. C., and Wolters, E. C. (2001). Loss of olfaction in de novo and treated Parkinson's disease : Possible implications for early diagnosis. *Movement Disorders*, 16(1) :6.

[Tolosa et al., 2006] Tolosa, E., Wenning, G., and Poewe, W. (2006). The diagnosis of Parkinson's disease. *The Lancet Neurology*, 5(1) :75–86.

[Tsanas et al., 2012a] Tsanas, A., Little, M. A., McSharry, P., and Ramig, L. O. (2012a). Using the cellular mobile telephone network to remotely monitor parkinsons disease symptom severity. *IEEE Transactions on Biomedical Engineering*.

[Tsanas et al., 2011] Tsanas, A., Little, M. A., McSharry, P. E., and Ramig, L. O. (2011). Nonlinear speech analysis algorithms mapped to a standard metric achieve clinically useful quantification of average Parkinson's disease symptom severity. *Journal of The Royal Society Interface*, 8(59) :842–855.

[Tsanas et al., 2012b] Tsanas, A., Little, M. A., McSharry, P. E., Spielman, J., and Ramig, L. O. (2012b). Novel Speech Signal Processing Algorithms for High-Accuracy Classification of Parkinson's Disease. *IEEE Transactions on Biomedical Engineering*, 59(5) :1264–1271.

[Vaiciukynas et al., 2017] Vaiciukynas, E., Verikas, A., Gelzinis, A., and Bacauskiene, M. (2017). Detecting Parkinson's disease from sustained phonation and speech signals. *PLOS ONE*, 12(10) :e0185613.

[Variani et al., 2014] Variani, E., Lei, X., McDermott, E., Moreno, I. L., and Gonzalez-Dominguez, J. (2014). Deep neural networks for small footprint text-dependent speaker verification. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4052–4056, Florence, Italy. IEEE.

[Viallet et al., 2010] Viallet, F., Gayraud, D., Bonnefoi, B., Renie, F., and Aurenty, R. (2010). Maladie de Parkinson idiopathique : aspects cliniques, diagnostiques et thérapeutiques. *EMC*.

[Viallet and Teston, 2007] Viallet, F. and Teston, B. (2007). La dysarthrie dans la maladie de Parkinson. In *Les Dysarthries*, pages 169–174. SOLAL.

[Vásquez-Correa et al., 2017a] Vásquez-Correa, J., Orozco-Arroyave, J. R., and Nöth, E. (2017a). Convolutional Neural Network to Model Articulation Impairments in Patients with Parkinson's Disease. In *INTERSPEECH*, pages 314–318. ISCA.

[Vásquez-Correa et al., 2017b] Vásquez-Correa, J. C., Serrà, J., Orozco-Arroyave, J. R., Vargas-Bonilla, J. F., and Nöth, E. (2017b). Effect of acoustic conditions on algorithms to detect Parkinson's disease from speech. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5065–5069.

[Weintraub et al., 2010] Weintraub, D., Koester, J., Potenza, M. N., Siderowf, A. D., Stacy, M., Voon, V., Whetteckey, J., Wunderlich, G. R., and Lang, A. E. (2010). Impulse Control Disorders in Parkinson Disease : A Cross-Sectional Study of 3090 Patients. *Archives of Neurology*, 67(5).

[Wu et al., 2018] Wu, K., Zhang, D., Lu, G., and Guo, Z. (2018). Influence of sampling rate on voice analysis for assessment of Parkinson's disease. *The Journal of the Acoustical Society of America*, 144(3) :1416–1423.

[Zhang et al., 2018] Zhang, H., Wang, A., Li, D., and Xu, W. (2018). DeepVoice : A voiceprint-based mobile health framework for Parkinson's disease identification. In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pages 214–217, Las Vegas, NV, USA. IEEE.

[Zhang, 2017] Zhang, Y. N. (2017). Can a Smartphone Diagnose Parkinson Disease ? A Deep Neural Network Method and Telediagnosis System Implementation.

Liste des publications

Publications :

- **L. Jeancolas**, G. Mangone, J.C. Corvol, M. Vidailhet, S. Lehéricy, B.E. Benkelfat, H. Benali, and D. Petrovska-Delacrétaz. "Comparison of Telephone Recordings and Professional Microphone Recordings for Early Detection of Parkinson's Disease, Using Mel-Frequency Cepstral Coefficients with Gaussian Mixture Models." In Interspeech 2019, 3033–37. ISCA, 2019. $https://doi.org/10.21437/Interspeech.2019-2825$.

- **L. Jeancolas**, G. Mangone, N. Villain, R. Gaurav, M.O. Habert, J.C. Corvol, M. Vidailhet, B.E. Benkelfat, S. Lehéricy, H. Benali, and D. Petrovska-Delacrétaz. "Analyse de La Voix Au Stade Débutant de La Maladie de Parkinson et Corrélations Avec Analyse Clinique et Neuroimagerie." In Journées d'Etude Sur La TéléSanté. Paris, France : Sorbonne Universités, 2019. $https://hal.archives-ouvertes.fr/hal-02161042$.

- **L. Jeancolas**, H. Benali, B.E. Benkelfat, G. Mangone, J.C. Corvol, M. Vidailhet, S. Lehéricy, and D. Petrovska-Delacrétaz. "Automatic Detection of Early Stages of Parkinson's Disease through Acoustic Voice Analysis with Mel-Frequency Cepstral Coefficients." In 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), 1-6, 2017. $https://doi.org/10.1109/ATSIP.2017.8075567$.

- **L. Jeancolas**, D. Petrovska-Delacrétaz, S. Lehéricy, H. Benali, and B.E. Benkelfat. "L'analyse de La Voix Comme Outil de Diagnostic Précoce de La Maladie de Parkinson : Etat de l'art." In CORESA 2016 : 18^{e} Edition COmpressions et REprésentation Des Signaux Audiovisuels, 113-21. Nancy : CNRS, 2016.
$https://projet.liris.cnrs.fr/coresa/articles/2016/Coresa2016-proceedings.pdf$.

Résumés présentés lors de conférences ou congrès :

- **L. Jeancolas**, G. Mangone, J.C. Corvol, M. Vidailhet, S. Lehéricy, B.E. Benkelfat, H. Benali, and D. Petrovska-Delacrétaz. "Parkinson's Disease Detection at an Early Stage Using Voice." Presented in 2^{nd} International Conference on Neurovascular & Neurodegenerative Disease (NVND) 2019, Paris, France. [oral]

- **L. Jeancolas**, N. Villain, R. Gaurav, D. Petrovska-Delacrétaz, B.E. Benkelfat, G. Mangone, M.O. Habert, J.C. Corvol, M. Vidailhet, S. Lehéricy, and H. Benali. "Voice Analysis in Early Parkinson's Disease and Correlation with Neuroimaging." Presented in 25^{th} Annual Meeting of Organization for Human Brain Mapping (OHBM) 2019, Rome, Italy. [poster]

- **L. Jeancolas**, D. Petrovska-Delacrétaz , H. Benali, and B.E. Benkelfat. “Analyse de La Voix Au Stade Débutant de La Maladie de Parkinson et Corrélats en Neuroimagerie”. Presented in 8^{th} Annual Meeting of Futur & Ruptures, 2019, Paris, France. [poster - best poster award]

- **L. Jeancolas**, D. Petrovska-Delacrétaz , H. Benali, and B.E. Benkelfat. “Diagnostic précoce de la maladie de Parkinson par l’analyse de la voix”. Presented in Colloque Evry Sciences et Innovation 2017, Evry, France. [poster - best poster award]

- **L. Jeancolas**, D. Petrovska-Delacrétaz , H. Benali, and B.E. Benkelfat. “Diagnostic précoce de la maladie de Parkinson par l’analyse de la voix”. Presented in SAMOVAR PhD day 2017, France. [poster - best poster award]

Annexe A : Protocole téléphonique

Exemple de ce que les sujets entendent au téléphone ...

... s'ils appellent à partir du numéro de téléphone qu'ils ont indiqué :

"Bonjour, merci pour votre appel, nous allons procéder à l'enregistrement de votre voix dans le cadre du protocole ICEBERG. Si vous prenez un traitement pour la maladie de Parkinson, tapez 1, sinon tapez 0."

1 : "Veuillez indiquer l'heure approximative de votre dernière prise et le nom du médicament en parlant après le BIP puis tapez 1."

"Très bien, nous allons procéder aux tâches vocales. Assurez-vous d'être si possible dans un endroit calme et veuillez ne pas utiliser le mode haut parleur, les écouteurs eux sont autorisés. La durée totale des tâches sera d'environ 15 min. Veuillez suivre les instructions audio au fur et à mesure, un bip marquera à chaque fois le début de l'enregistrement. A tout moment si vous souhaitez refaire la tâche vocale en cours tapez 0. Tapez 1 quand vous êtes prêt."

1 : "Veuillez dire le son « aaa » le plus longtemps possible sans respirer en gardant une voix constante. Ex « aaa ». Quand vous aurez fini tapez 1. C'est à vous."

"Maintenant veuillez dire le son « aaa » sans respirer en faisant varier la hauteur de la voix, à la manière d'une sirène. Vous commencerez par les graves pour aller dans les aigus et redescendre dans les graves. Ex « aaa ». Quand vous aurez fini tapez 1. C'est à vous."

"Veuillez répéter aussi rapidement et longtemps que possible la syllabe /pa/ sans respirer. Ex : « papapapapa ». Quand vous aurez fini tapez 1. C'est à vous."

"Veuillez faire la même chose avec la syllabe /pou/, toujours sans respirer.
Ex : « poupoupou ». Quand vous aurez fini tapez 1. C'est à vous."

"Veuillez faire la même chose avec les syllabes /poupa/, toujours sans respirer.
Ex « poupapoupa ». Quand vous aurez fini tapez 1. C'est à vous."

"Veuillez faire la même chose avec la syllabe /kou/, toujours sans respirer.
Ex « koukoukou ». Quand vous aurez fini tapez 1. C'est à vous."

"Veuillez faire la même chose avec les syllabes /pakou/, toujours sans respirer.
Ex « pakoupakou ». Quand vous aurez fini tapez 1. C'est à vous."

"Veuillez faire la même chose avec les syllabes /pataka/, toujours sans respirer.
Ex « patakapataka ». Quand vous aurez fini tapez 1. C'est à vous."

"Veuillez répéter les phrases suivantes après le BIP puis terminez en tapant 1 :
Tu as appris la nouvelle ? BIP
Tu as bien raison ! BIP
C'est pas possible ! BIP
Comment il s'appelle déjà ? BIP
Tu sais ce qu'il est devenu ? BIP
Il n'aurait jamais du faire ça ! BIP
Les chiens aiment courir après les ballons. BIP
Un carré est un rectangle particulier. BIP."

"Veuillez maintenant nous raconter votre journée ou celle d'hier pendant environ 1 min. L'enregistrement commencera au 1er bip et un message sonore vous indiquera quand cela fera 1min. C'est à vous."

"Cela fait 1 min merci" "Maintenant en respirant quand vous voulez, veuillez répéter la syllabe /pa/ en suivant le rythme que vous allez entendre dans l'exemple, et ce pendant 30 sec. Un message sonore vous indiquera quand cela fera 30s. Ex : « pa pa pa ». C'est à vous."

"Cela fait 30s merci" "Maintenant veuillez faire la même chose avec la syllabe /kou/.
Ex « kou kou kou ». C'est à vous."

"Cela fait 30s merci" "Maintenant veuillez faire la même chose avec les syllabes /pa kou/.
Ex «pa kou pa kou». C'est à vous."

"Cela fait 30s merci" "Pour finir voici une citation de ... à répéter après le BIP . Quand vous aurez fini tapez 1 : ... BIP"

"L'enregistrement est terminé, merci pour votre appel, si vous avez une question ou un commentaire vous pouvez nous laisser un message après le bip. Si durant cette cession vous avez utilisé des écouteurs merci de nous le préciser en commentaire. Sinon nous vous disons à bientôt. Au revoir."

... s'ils appellent à partir d'un autre téléphone que celui qu'ils ont renseigné :

"Bonjour, nous n'avons pas reconnu votre numéro de téléphone. Si vous appelez pour un enregistrement mensuel de la voix dans le cadre du protocole ICEBERG, merci de nous appeler à partir du numéro que vous nous avez donné. (En cas de problème ou de changement de numéro vous pouvez joindre Laetitia Jeancolas au 06XXXXXXXX). Merci et à bientôt."

SMS que les sujets concernés reçoivent chaque mois :
Bonjour, merci de nous appeler au 01XXXXXXXX afin d'effectuer votre enregistrement téléphonique mensuel de la voix, dans le cadre du protocole ICEBERG. Vous avez 3 jours à partir de maintenant pour appeler au moment qui vous convient le mieux. La durée de l'appel sera d'environ 15 min. A tout moment pour réécouter une consigne ou refaire la tâche en cours vous pouvez taper 0. Généralement il vous sera demandé de taper 1 pour passer à la tâche suivante, sauf pour les dernières tâches où la fin sera précisée par un message sonore. En cas de problème vous pouvez joindre Laetitia Jeancolas au 06XXXXXXXX. Merci beaucoup et à bientôt.

Annexe B : Mise en place du répondeur interactif

Le répondeur interactif

Nous avons écarté les répondeurs analogiques à cause de leur coût et du fait qu'ils ne sont quasiment plus utilisés. Nous nous sommes donc tournés vers les possibilités de répondeur VoIP et avons opté pour le logiciel IP IVM de l'entreprise NCH pour sa facilité de programmation. Il existe un logiciel libre Asterisk, gratuit, mais qui est d'une programmation plus basique et aurait engendré un délai de mise en œuvre supplémentaire pour la mise en place du serveur téléphonique. Nous avons installé IVM sur un ordinateur portable nous servant de serveur téléphonique. Nous avons programmé IVM de manière à ce qu'il puisse jouer les consignes quand on l'appelle et enregistrer l'exécution des tâches, associées au numéro de téléphone de l'appelant, sur le serveur. Pour passer d'une tâche à l'autre, le participant appuie sur une certaine touche, il peut appuyer sur une autre touche s'il souhaite refaire la tâche.

Ligne SIP

Le choix de la ligne qui allait transmettre le son de l'appelant jusqu'au répondeur a été assez complexe. Nous avons mis de côté les lignes de téléphonie classique (qui auraient pu fonctionner avec un modem voix) car vu le nombre de participants qui allaient appeler tous les mois, on voulait pouvoir avoir au moins 2 appels simultanés, ce qui n'était possible qu'avec les lignes SIP (de VoiP).

Dans un premier temps, nous avons testé les lignes SIP de Télécom SudParis, nous n'avons pas retenu cette possibilité entre autres car les appels simultanés n'étaient pas possibles.

Nous nous sommes alors tournés vers la ligne SIP de ma Freebox personnelle, qui avait l'option d'appels simultanés. Nous avons rencontré un problème avec la transmission de l'information généré par l'appui sur les touches (nécessaire pour passer à la tâche suivante). En effet le codage de l'information de pression des touches se fait en téléphonie classique grâce au code DTMF : la pression des touches est transformée en un son de fréquence particulière qui est transmis par le même canal que les autres sons. Pour le cas de la ligne SIP de free, il utilise le protocole SIP Info qui va laisser le décodage du son au répondeur. Or ce protocole est mal supporté par IVM, conduisant souvent au non décodage d'envois DTMF. Depuis, la ligne SIP de Free a été retirée du commerce.

Orange propose aussi des lignes SIP mais il s'avère que les identifiants SIP sont masqués et non accessibles, ne pouvant donc être utilisés avec des répondeurs VoIP.

Nous avons ensuite loué une ligne OVH (grand fournisseur de ligne SIP). Après plusieurs tests nous nous sommes aperçus qu'assez régulièrement il y avait des problèmes avec les consignes audio se manifestant soit par un raccroché à la lecture des fichiers audio, soit par une mauvaise qualité audio. Après investigation auprès d'OVH, ces derniers nous ont conseillé d'utiliser le codec G.729 pour résoudre ces problèmes de qualité, codec malheureusement non supporté par IVM.

Nous nous sommes donc tournés vers un autre fournisseur de ligne SIP : ippi, qui s'est avéré fonctionner mieux avec le codec G.711 et n'utilisant pas le protocole SIP Info pour l'envoi des codes DTMF. Cependant des problèmes de connexion persistaient nous poussant à investiguer du côté de la connexion internet utilisée par notre serveur téléphonique.

Connexion internet

Nous avions dans un premier temps connecté le serveur téléphonique au réseau internet de l'hôpital. L'hôpital attribuait aléatoirement au serveur un port qui se retrouvait de temps en temps bloqué par le par-feu de l'hôpital, empêchant la connexion avec la ligne SIP .

Nous avons donc loué une box (la livebox pro V4 d'orange) de manière à avoir un réseau internet sans port bloqué. Après deux changements de box nous avons fini par obtenir un réseau internet stable et sans problème de port.

Restait à vérifier si le débit de la connexion internet était suffisant pour garantir une bonne qualité, notamment lors d'appels simultanés. Le débit disponible avec notre connexion internet (utilisant l'ADSL) est de 900 kbit/s en débit montant et de 7Mbit/s en débit descendant.

Nous avons mesuré le débit utilisé lors d'un appel et avons trouvé une moyenne d'environ 100 kbit/s en débit montant (consignes audio) et en débit descendant (voix de l'appelant). Le débit disponible semblait largement suffisant pour traiter 2 appels simultanés. On pouvait néanmoins rencontrer un problème de débit quand on utilisait le serveur, à distance, via un outil graphique comme Team Viewer (ce qui était nécessaire pour l'envoi régulier des sms de rappel, ou pour l'ajout des numéros des nouveaux participants dans le logiciel). Les débits montants et descendants pouvaient alors atteindre 100% de la capacité. Même si le protocole RTP est prioritaire, lorsqu'on utilise Team Viewer en même temps qu'un sujet appelle, cela pouvait nuire à la qualité des consignes et des enregistrements. Orange nous a proposé de passer notre débit descendant disponible de 8 à 20Mbit/s, mais ne pouvait augmenter notre débit montant car il aurait fallu passer en VDSL ou à la fibre et le bâtiment n'était éligible à aucun des deux. Nous avons rencontré de nombreux problèmes de connexion suite à l'augmentation à 20Mbit/s. Nous sommes alors retournés au débit de 8Mbit/s et évitons d'ouvrir une session de connexion à distance quand un participant est en train d'effectuer les tâches vocales.

Le dernier point indispensable au bon fonctionnement du serveur téléphonique a été de brancher le serveur et la box sur une prise alimentée par un réseau ondulé pour ne pas subir les coupures de courant mensuelles servant à la maintenance électrique de l'ICM (endroit où est installé le serveur).

Suivi des appels

De manière à pouvoir relier les participants à leurs enregistrements, ces derniers sont identifiés par leur numéro de téléphone quand ils appellent (leur numéro étant rentré préalablement dans le logiciel du répondeur). S'ils appellent d'un autre numéro, un message d'erreur les informe qu'ils doivent appeler à partir du numéro qu'ils nous ont communiqué. Une fois reconnus, ils sont guidés par le répondeur pour effectuer les différentes tâches vocales. Chaque tâche effectuée est enregistrée dans un fichier wav, avec comme information le numéro de téléphone, la date, le nom de la tâche et un numéro attribué à l'appel (afin de pouvoir discerner 2 sessions qui seraient effectuées le même jour). Pour l'analyse, les numéros de téléphones sont remplacés par le numéro d'identification du sujet.

Afin d'éviter que trop de sujets appellent en même temps (nous avions opté pour 2 appels simultanés possibles), nous avons essayé de répartir les appels des 200 sujets sur le mois. Nous

envoyons tous les 3 jours un sms (ou un mail suivant la préférence des participants) à une vingtaine de sujets les invitant à procéder à leur enregistrement mensuel. Si nous voyons qu'ils n'ont pas fait leur enregistrement dans les 5 jours, un sms de rappel leur est envoyé, sachant qu'à chaque fois mon numéro de portable leur est donné pour qu'ils me préviennent en cas de problème, de question, ou d'indisponibilité pour faire les enregistrements. Comme envoyer tous les 3 jours les sms, vérifier à chaque fois qui a fait son enregistrement et envoyer les rappels, prendrait beaucoup trop de temps si on le faisait à la main, nous avons conçu des scripts VBScript et VBA pour semi automatiser ce suivi et générer la liste des numéros de téléphones ou des adresses mail à qui il faut envoyer le sms ou le mail, ainsi que la liste des numéros pour le message de rappel. Ensuite l'option sms illimités à laquelle nous avons souscrite avec la livebox, nous permet d'envoyer les sms à partir de la messagerie mail d'orange à condition qu'elle soit utilisée par un ordinateur connecté à la box. Or comme je ne travaille pas au même endroit que là où est le serveur téléphonique, j'utilise un outil graphique de connexion à distance pour envoyer ces sms, en évitant les moments où des sessions d'enregistrements sont en cours.

Pour savoir si une session d'enregistrement est en cours et pour suivre le déroulement des appels, un autre script détecte les appels sur le serveur téléphonique et m'envoie aussitôt un sms. Un autre sms m'est envoyé quand la session est arrivée à son terme. De même un sms m'est envoyé automatiquement quand un problème de connexion important est détecté sur le serveur, pour que je puisse réagir vite et éviter que trop de participants soient confrontés à un serveur non fonctionnel.

Titre : Détection précoce de la maladie de Parkinson par l'analyse de la voix et corrélations avec la neuroimagerie

Mots clés : Maladie de Parkinson, Analyse de la voix, Traitement du signal, Apprentissage supervisé, Modèles de prédiction, Télédiagnostic

Résumé : Les modifications de la voix, prenant la forme de dysarthrie hypokinétique, sont un des premiers symptômes à apparaître dans la maladie de Parkinson (MP). Un grand nombre de publications existent sur la détection de MP par l'analyse de la voix, mais peu se sont intéressées spécifiquement au stade débutant. D'autre part, à notre connaissance, aucune étude n'avait été publiée sur la détection de MP via des enregistrements issus du réseau téléphonique. L'objectif de cette thèse a été d'étudier les modifications de la voix aux stades débutant et préclinique de la maladie de Parkinson, et de développer des modèles de détection précoce automatique et de suivi de cette maladie. Le but à long terme étant de pouvoir construire un outil de diagnostic précoce et de suivi, peu couteux, utilisable par les médecins en cabinet, et de manière encore plus intéressante, à partir de n'importe quel téléphone.
La première étape a été de constituer une grande base de données voix de plus de 200 locuteurs français, comprenant des sujets MP débutants, des sujets sains et des sujets atteints de trouble idiopathique du comportement en sommeil paradoxal (iRBD), pouvant être considérés comme au stade préclinique de la maladie de Parkinson. Les participants ont effectué différentes tâches vocales enregistrées avec un microphone professionnel et avec le microphone interne d'un ordinateur. De plus, une fois par mois, ils ont également effectué ces tâches en appelant un servant vocal interactif à partir de leur propre téléphone. Nous avons étudié les effets de la qualité des microphones, du type de tâches, du genre, et de la méthode de classification.
Nous avons analysé ces enregistrements vocaux par le biais de trois méthodes d'analyses différentes, couvrant différentes échelles de temps. Nous avons commencé avec des coefficients cepstraux et des modèles de mélange gaussien (GMM). Ensuite nous avons adapté la méthodologie des x-vecteurs (qui n'avait jamais été utilisée pour la détection de MP), puis nous avons extrait des paramètres globaux que nous avons classés avec des machines à vecteurs de support (SVM). Nous avons constaté des perturbations vocales aux stades débutant et préclinique de MP dans plusieurs domaines phonétiques, tels que l'articulation, la prosodie, la fluence verbale et les capacités rythmiques.
Avec les enregistrements du microphone professionnel, nous sommes parvenus à détecter les hommes MP débutants avec une précision (Acc) de 89%, à partir de 6min de lecture, monologue et répétitions rapides et lentes de syllabes. Concernant les femmes, nous avons atteint Acc=70% à partir d'1min de monologue. Avec les enregistrements téléphoniques, nous avons obtenu des performances de classification de 75% pour les hommes, à partir de 5min de répétitions rapides de syllabes, et de 67% pour les femmes, à partir de 5min de monologue. Ces résultats constituent un premier pas important vers un télédiagnostic précoce de la maladie de Parkinson.
Enfin nous avons aussi étudié les corrélations avec les données de neuroimagerie. Nous avons pu prédire linéairement, de manière significative, les données de DatScan et d'imagerie par résonance magnétique (IRM) sensible à la neuromélanine, à partir de paramètres vocaux. Ce résultat est prometteur au vu d'une possible utilisation future de la voix pour le suivi de l'évolution des premiers stades de MP.

université PARIS-SACLAY
ÉCOLE DOCTORALE
Sciences et technologies de l'information et de la communication (STIC)

Title : Early detection of Parkinson's disease using voice analysis and correlations with neuroimaging

Keywords : Parkinson's disease, Voice analysis, Signal processing, Supervised learning, Predictive modeling, Telediagnosis

Abstract : Vocal impairments, known as hypokinetic dysarthria, are one of the first symptoms to appear in Parkinson's Disease (PD). A large number of articles exist on PD detection through voice analysis, but few have focused on the early stages of the disease. Furthermore, to our knowledge, no study had been published on remote PD detection via speech transmitted through the telephone channel. The aim of this PhD work was to study vocal changes in PD at early and preclinical stages, and develop automatic detection and monitoring models. The long-term purpose is to build a cheap early diagnosis and monitoring tool, that doctors could use at their office, and even more interestingly, that could be used remotely with any telephone.

The first step was to build a large voice database with more than 200 French speakers, including early PD patients, healthy controls and idiopathic Rapid eye movement sleep Behavior Disorder (iRBD) subjects, who can be considered at PD preclinical stage. All these subjects performed different vocal tasks and were recorded with a professional microphone and with the internal microphone of a computer. Moreover, they called once a month an interactive voice server, with their own phone. We studied the effect of microphone quality, speech tasks, gender, and classification analysis methodologies.

We analyzed the vocal recordings with three different analysis methods, covering different time scale analyses. We started with cepstral coefficients and Gaussian Mixture Models (GMM). Then we adapted x-vectors methodology (which never had been used in PD detection) and finally we extracted global features classified with Support Vector Machine (SVM). We detected vocal impairments at PD early and preclinical stages in articulation, prosody, speech flow and rhythmic abilities.

With the professional microphone recordings, we obtained an accuracy (Acc) of 89% for male early PD detection, just using 6min of reading, free speech, fast and slow syllable repetitions. As for women, we reached Acc = 70% with 1min of free speech. With the telephone recordings, we achieved Acc = 75% for men, with 5min of rapid syllable repetitions, and 67% for women, with 5min of free speech. These results are an important first step towards early PD telediagnosis.

We also studied correlations with neuroimaging, and we were able to linearly predict DatScan and Magnetic Resonance Imaging (MRI) neuromelanin sensitive data, from a set of vocal features, in a significant way. This latter result is promising regarding the possible future use of voice for early PD monitoring.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France