Caractérisation automatique du rythme de la parole : application aux cancers des voies aéro-digestives supérieures et à la maladie de Parkinson

Robin Vaysse

▶ **To cite this version:**

Robin Vaysse. Caractérisation automatique du rythme de la parole : application aux cancers des voies aéro-digestives supérieures et à la maladie de Parkinson. Sciences de l'information et de la communication. Université Paul Sabatier - Toulouse III, 2023. Français. NNT : 2023TOU30062 . tel-04198849

HAL Id: tel-04198849
https://theses.hal.science/tel-04198849

Submitted on 7 Sep 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Fédérale
Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Robin VAYSSE

Le 21 mars 2023

Caractérisation automatique du rythme de la parole : application aux cancers des voies aéro-digestives supérieures et à la maladie de Parkinson

---

Ecole doctorale : **EDMITT - Ecole Doctorale Mathématiques, Informatique et Télécommunications de Toulouse**

Spécialité : **Informatique et Télécommunications**

Unité de recherche :
IRIT : Institut de Recherche en Informatique de Toulouse

Thèse dirigée par
Jérôme FARINAS et Corine ASTESANO

Jury

M. François PELLEGRINO, Rapporteur
Mme Elisabeth DELAIS-ROUSSARIE, Rapporteure
Mme Virginie WOISARD-BASSOLS, Examinatrice
M. Jérôme FARINAS, Directeur de thèse
Mme Corine ASTESANO, Co-directrice de thèse
Mme Cécile FOUGERON, Présidente

Remerciements

Tout d'abord, je tiens à remercier grandement mes directeurs de thèse : Corine Astésano et Jérôme Farinas qui m'ont fait confiance et m'ont accompagné durant ces trois années de travail. Ils ont été des soutiens scientifiques mais aussi psychologiques notamment lors des périodes compliquées des confinements. Merci Corine pour m'avoir guidé dans le domaine des sciences du langage qui m'était inconnu jusqu'alors. Je prêcherai désormais en faveur du Saint 600ms. Merci Jérôme pour la bienveillance que tu as eue envers moi ainsi que pour tes conseils toujours pertinents lors de nos réunions.

Merci également à tous mes collègues de l'équipe SAMoVA. Travailler dans cette équipe a été un réel plaisir. Merci aux permanents de l'équipe qui ont pu m'aider de près ou de loin durant mes travaux : Isabelle, Régine, Hervé, Thomas, Julie, Christine. Mention spéciale pour notre chef d'équipe aveyronnais Julien qui a toujours su se rendre disponible, cela malgré un emploi du temps nécessitant des capacités de clonage. Merci aussi à tous les amis avec qui j'ai passé de très bons moments à l'IRIT et en dehors : Léo pour son hospitalité, Lucile avec qui j'ai pu partager ma première conférence à Brno, Tim pour sa gentillesse sans limite, Sebastião grâce à qui j'ai acquis un accent portugais irréprochable, Jim pour son humour et ses talents de musicien, Vincent pour tous tes conseils de vie et de thèse et Lila avec qui j'ai passé de très bons moments durant notre collocation dans le bureau 227. Merci également à toutes les personnes avec qui j'ai pu échanger ou jouer au tarot le midi : Étienne, Alexis, Romain, Mathieu, Benjamin et bien d'autres...

Je remercie évidemment grandement Verdiana avec qui j'ai partagé un bureau pendant plus de 2 ans. Sa bonne humeur contagieuse et nos discussions (parfois futiles) pendant nos pauses ont fortement contribué à rendre mes années de thèses agréables même dans les moments les plus difficiles.

Je remercie l'ensemble des membres du projet RUGBI qui m'ont aidé à orienter mes recherches lors de nos réunions.

Merci à l'ensemble des mes amis qui me permettent tout simplement d'être moi-même : Adil, Geoffrey, Max et Raphaël grâce à qui j'apprends de nombreuses choses et avec qui j'ai voyagé et passé des moments inoubliables. William avec qui nous nous sommes qualifiés deux fois pour les phases finales du concours IDAO [1] bien que la chance a fait que nous n'avons jamais pu assister physiquement aux phases finales. Les Mapi3 : Arthur, Hugo, Quentin (et Alexis même si tu n'es plus sur Toulouse) qui étaient toujours disponibles pour boire une (ou plusieurs) bières en fin de semaine.

Un merci particulier à Anaël avec qui j'ai passé des moments inoubliables de la maternelle à aujourd'hui. Bien qu'on ne se voie pas aussi souvent qu'on le voudrait, je sais qu'on pourra toujours compter l'un sur l'autre.

Je remercie ma famille qui m'a toujours soutenu dans mes choix et grâce à qui je pouvais échapper à tous mes soucis le temps d'un week-end dans ma campagne

1. https://idao.world/

Aveyronnaise natale.

Enfin, merci à Judith qui me soutient et me supporte depuis plus de 10 ans maintenant. Bien que ces derniers mois de thèse n'aient pas été faciles, elle m'a permis de tenir le coup. Nous avons et nous continuerons d'évoluer ensemble à l'avenir vers de nouveaux sommets (dans la vie et en escalade bien sûr !).

Table des matières

Table des figures

Glossaire

C2SI : *Carcinologic Speech Severity Index*, projet financé par l'Institut National du Cancer.

Consensus de Delphes : Le consensus de Delphes est une méthode de recherche qui vise à obtenir l'opinion d'un groupe d'experts sur un sujet précis en utilisant une série de questions et de réponses.

$f0$: Fréquence fondamentale.

LASSO : *Least Absolute Shrinkage and Selection Operator*.Méthode de réduction des coefficients de régression statistique.

MDP : Maladie de Parkinson.

ORL : Oto-rhino-laryngologie.

RUGBI : Projet financé par l'Agence Nationale de la Recherche.

SVM : *Support Vector Machine*, Machine à Vecteur de Support. Algorithme d'apprentissage automatique pour la classification.

SVM : *Support Vector Regressor*. Algorithme d'apprentissage automatique pour la régression adapté de l'algorithme SVM.

TSNE : *T-distributed Stochastic Neighbor Embedding*. Algorithme de réduction de dimensions.

VADS : Voies aéro-digestives supérieures.

Introduction

La modélisation de la prosodie

La prosodie est un élément essentiel de la parole. Elle constitue un moyen de transmettre des informations linguistiques et extra-linguistiques, telles que l'emphase, le sens, la structure du discours ou les émotions. Elle joue un rôle important dans l'identification et la reconnaissance des mots. L'un des buts principaux de la prosodie est de segmenter les énoncés de parole en unités linguistiques plus courtes et de les organiser de manière cohérente pour l'auditeur. Ainsi, trois principes organisateurs sont mis en jeux dans la prosodie : l'intonation, l'accentuation et le rythme. Ces trois processus s'organisent autour d'un centre commun qui est la métrique. En effet, la métrique organise l'accentuation et pose les règles du rythme dans une langue donnée tandis que l'intonation modélisée par les variations de fréquence fondamentale vient se poser sur les différents accents produits lors de la parole. Les groupes rythmiques (ou « chunks ») constituent les groupements de bas niveau, domaines de l'accentuation. Les groupes intonatifs, à l'interface avec la syntaxe et les contraintes sémantico-pragmatiques, se situent aux niveaux supérieurs de l'organisation prosodique.

La métrique impose donc les règles de l'accentuation. Ainsi, une syllabe accentuée ne peut pas se situer n'importe où. Le français par exemple possède une double accentuation initiale et finale (Di Cristo, 1999) qui implique une accentuation potentielle de la première et dernière syllabe d'un mot ou d'un groupe de mots. Ces accents sont alors organisés hiérarchiquement de sorte à structurer les énoncés.

Le rythme de la parole est alors caractérisé par cette alternance de syllabes fortes et faibles au travers des accents métriques. Il existe de nombreuses théories formelles de la métrique et du rythme de la parole. Bien que les modèles théoriques soient encore discutés et éprouvés, de nombreuses modélisations automatiques proposent de décrire le rythme de la parole. Cependant, il n'existe pas, à l'heure actuelle, de modélisation du rythme qui puisse rendre compte efficacement de toutes les dimensions du rythme de façon automatique, même si certaines méthodes se rapprochent de la vision linguistique du rythme. Nous pouvons par exemple citer les modélisations basées sur l'étude des modulations d'amplitudes (Tilsen et Johnson, 2008; Liss et collab., 2010; Gibbon, 2021) et de fréquences (Sheft et collab., 2008, 2012; Varnet et collab., 2017; Gibbon, 2021) de la parole qui étudient les variations lentes des éléments qui composent le signal de parole. Ces modèles permettent de repérer les régularités de la

parole à différents niveaux. Ces modélisations du rythme pourraient donc faciliter la détection des troubles prosodiques chez des individus en caractérisant les structures rythmiques de leurs énoncés. En effet, une bonne structuration prosodique permet d'améliorer l'accès au sens des auditeurs. Il est donc crucial de pouvoir quantifier ces déficits.

Le cas particulier de la parole pathologique

Dans ce travail de doctorat, nous nous intéressons à l'application de ces modélisations automatiques du rythme de la parole à l'étude de la parole pathologique. Ce travail s'inscrit dans le cadre du projet ANR RUGBI[2] « *Recherche d'unités linguistiques pertinentes pour améliorer la mesure de l'intelligibilité de la parole altérée par des troubles de production pathologique* ». Ce projet a pour but de participer à l'amélioration de la mesure des déficits d'intelligibilité de la parole. Dans le cadre de pathologies comme les cancers de la bouche ou de l'oropharynx (cancer des Voies Aéro-Digestives Supérieures : VADS), des traitements médicaux lourds comme des chimiothérapies et/ou des chirurgies sont généralement mis en place. À la suite de cela, il est courant que les patients présentent des troubles de la communication orale. Ces troubles sont variables et diffèrents d'un patient à un autre, en impactant leurs productions orales de multiples façons. Ainsi, la qualité de vie des patients peut diminuer suite à un déficit dans leurs capacités à transmettre des messages par la voix (Guibert et collab. (2011)). Cette difficulté à transmettre leurs intentions à un interlocuteur impacte négativement leur vie professionnelle et sociale. Ces patients sont alors pris en charge par des spécialistes (médecins ORL, orthophonistes) qui évaluent régulièrement leurs troubles dans le cadre du suivi de leurs traitements. Cependant, les spécialistes de la parole pathologique, en suivant régulièrement leurs patients, peuvent s'adapter à leur production et induire un phénomène d'habituation, qui peut rendre difficile l'évaluation objective de la qualité de la voix à des fins thérapeutiques. Il est donc nécessaire de procéder à des évaluation avec de multiples experts pour obtenir une évaluation objective sur de la parole lue ou spontanée (par exemple, description d'image ; Dittner et collab. (2010); Woisard et Lepage (2010); Balaguer et collab. (2019a,b)). Malheureusement, c'est une méthode qui induit une latence assez grande entre la passation et les résultats, et qui mobilise beaucoup de ressources pour l'évaluation. Ceci n'est donc pas idéal dans le cadre de suivi de patients lors de consultations. De plus, une évaluation de ce type mène à une évaluation globale des troubles, sans identifier leur source exacte.

L'intérêt du projet RUGBI est donc de chercher des solutions à ces limitations, en proposant des outils automatiques d'évaluation afin d'aider les médecins dans la mesure et la localisation des troubles des patients. Les mesures plus fines proposées dans le cadre du projet RUGBI permettent de mettre en évidence les caractéristiques qui induisent une perte d'intelligibilité, ainsi que leur localisation précise. Le développement de ce genre d'outils pourrait alors permettre de suivre plus efficacement les patients et comprendre les spécificités de leurs troubles. Cela faciliterait également le

2. ANR-18-CE45-0008

suivi des patients afin de vérifier l'efficacité de leur rééducation. Pour cela, le projet se base sur deux corpus de parole pathologique. Le premier est composé de patients atteints de cancers des Voies Aéro-Digestives Supérieures (VADS). Cette pathologie et les traitements qu'elle induit favorise les troubles périphériques d'ordre respiratoires et de la déglutition. La prosodie de ces personnes est donc logiquement impactée de par ces difficultés qui entraînent une augmentation des pauses respiratoires ainsi qu'une baisse du débit de parole. Les cancers localisés au niveau de la langue ou du palais vont également jouer un rôle dans l'articulation des patients. La production de syllabes intelligibles devenant plus difficile, il est alors intéressant de voir quels mécanismes compensatoires les patients peuvent mettre en place pour améliorer leur production. En plus de patients atteints de cancers VADS, le projet RUGBI intègre un ensemble d'enregistrements de personnes atteintes de la Maladie de Parkinson (MDP). Cette maladie neurologique provoque des troubles moteurs centraux. L'impact de cette pathologie sur la parole n'est pas systématique à un stade précoce. Néanmoins, parmi les altérations connues à ce jour, nous pouvons en citer deux types : les troubles laryngés entraînants des variations anormales de la fréquence fondamentale, et les disfluences avec l'apparition de courts segments de parole, avec un débit variable et des silences mal placés (Darley et collab., 1969a,b; Logemann et collab., 1978).

Sur la base de ces observations, nous considérons que la prosodie est l'un des axes majeurs à explorer dans l'étude des symptômes de ces maladies. Parmi les paramètres prosodiques, le rythme est le socle de la structuration de la parole. Si la structuration rythmique est préservée, les groupements prosodiques sont cohérents et l'accès au sens facilitée pour l'auditeur.

Objectifs de cette thèse

Les objectifs de mes travaux sont donc multiples. Premièrement, il est crucial de trouver une modélisation automatique du rythme qui soit capable de prendre en compte les différentes caractéristiques du rythme linguistique. Cela sous-entend de définir ce qu'est le rythme de la parole et s'assurer que le choix ou la création d'une modélisation du rythme soit fidèle à cette définition. Une fois la cohérence de cette modélisation validée par une comparaison avec des analyses prosodiques manuelles, nous l'appliquons à la parole pathologique. Ainsi, dans le cadre du projet RUGBI, nous proposons d'extraire automatiquement un ensemble de caractéristiques du rythme. Ces caractéristiques sont ensuite utilisées afin de vérifier si le rythme peut apporter de l'information pertinente pour la modélisation d'un score d'intelligibilité. Enfin, via ces caractéristiques, nous espérons pouvoir caractériser au mieux les différentes stratégies employées par les patients pour compenser leurs troubles de la parole. La spécification précise des troubles rythmiques pourrait donc permettre à terme de proposer des pistes de travail pour la rééducation des patients.

Ce manuscrit est composé de cinq chapitres. Le chapitre 1 présente un état de l'art dans lequel nous posons les fondements linguistiques du rythme sur lesquels nous nous sommes basés tout au long de cette thèse. Nous nous intéresserons donc

à la structuration prosodique du français, ainsi qu'aux manifestations physiques et perceptives du rythme de la parole. Nous présentons par la suite dans le chapitre 2 l'évolution des méthodes automatiques couramment utilisées pour étudier le rythme, en montrant les liens existants entre ces modélisations automatiques et les modèles linguistiques évoqués précédemment. Pour cela, nous réalisons un historique des méthodes qui ont été utilisées pour modéliser le rythme dans différents domaines, dont l'identification des langues et la recherche automatique du tempo en musique.

Dans le chapitre 3, nous présentons en détail les corpus de parole que nous avons étudiés dans nos travaux. Parmi eux, nous exposerons un petit corpus de slam qui nous sera utile pour éprouver les modélisations automatiques sélectionnées. Le slam étant fortement rythmique (régularité métrique) par nature, nous voulions tester nos modélisations afin de vérifier qu'elles rendent bien compte des régularités rythmiques de la parole. Nous présentons également les deux corpus de parole pathologiques qui composent le projet RUGBI. Ces corpus contiennent des échantillons de parole saine, de patient atteints de cancers VADS et de patients atteints de la MDP. En plus des enregistrements audio, ces corpus sont composés de méta-données et annotations prosodiques. Parmi elles, nous avons des annotations cliniques produites par des médecins afin de quantifier les troubles généraux de la parole. En plus de celles-ci, nous avons produit des annotations prosodiques de certains enregistrements que nous pourrons comparer aux résultats de nos modélisations.

Le chapitre 4 est dédiée à la mise en place des modélisations que nous avons choisi de développer. Afin d'éprouver ces méthodes, nous les appliquons à notre corpus de slam. Nous vérifions ainsi que l'ensemble des caractéristiques rythmiques du slam est détecté par ces modèles. Nous détaillons alors comment nous avons adapté les méthodes existantes avec les différentes étapes de raffinement pour pouvoir décrire de façon pertinente le rythme de la parole.

Enfin, le chapitre 5 détaille les diverses analyses et paramètres que nous avons pu extraire de nos modélisations. Nous appliquons les méthodes sélectionnées à la parole pathologique, afin d'étudier les particularités prosodiques des différentes pathologies. À partir de nos observations, nous pouvons ensuite extraire un ensemble de caractéristiques rythmiques automatiquement. Nous nous servons de ces paramètres dans le but de modéliser l'intelligibilité des locuteurs à l'aide d'algorithme d'apprentissage automatique. En plus d'une modélisation globale de l'intelligibilité, nous testons des caractéristiques automatiques en les comparant à des analyses perceptives de la parole. Cette comparaison nous apporte des éléments de compréhension sur les stratégies prosodiques des patients, et nous permet de dégager des regroupements en fonction des particularités rythmiques des locuteurs.

1

Les fondements du rythme de la parole

Sommaire

La prosodie de la parole est le processus qui gère les aspects mélodiques, intonatifs et rythmiques de la parole. Elle permet de comprendre comment les variations de hauteur, de durée et d'intensité de la voix influencent le sens et la signification des mots et des phrases. La prosodie est un élément crucial de la communication verbale, car elle peut influencer la perception des émotions, de l'attitude et de l'intention de l'émetteur. Elle permet au locuteur de transmettre des informations qui ne sont pas contenues dans le sens des mots eux-mêmes. Au delà de ces aspects sémantico-pragmatiques, la prosodie permet également d'organiser le flux de parole, tant pour le locuteur que pour l'auditeur, via un découpage du signal en éléments organisés hiérarchiquement. Concrètement, la prosodie est la manifestation de plusieurs paramètres, à savoir : l'accentuation qui marque localement des syllabes accentuées sur des mots lexicaux, l'intonation qui se manifeste par les variations de fréquence fondamentale (f_0) globales, les pauses et enfin le rythme de la parole. Dans le modèle de l'approche métrique

(Liberman et Prince, 1977; Halle et Vergnaud, 1987) adaptée au français par Di Cristo (2000), le rythme est considéré comme un élément fondateur de la prosodie. Ainsi, le rythme se manifeste au travers de l'accentuation via l'alternance entre les syllabes accentuées et les syllabes inaccentuées. L'intonation viendrait alors "se poser" sur les variations rythmiques. Ce chapitre sera donc dans un premier temps consacré à l'étude de l'accentuation et de son organisation dans le cadre du français. Par la suite, des éléments de définitions sur le rythme de la parole seront apportés avant de décrire comment se manifeste le rythme du français par rapport à d'autres langues.

1.1 Prosodie et structuration hiérarchique de la parole

L'accentuation de la parole joue un rôle majeur dans la description du rythme parolier. Il est donc indispensable d'introduire les fondements de l'accentuation avant de pouvoir aborder la notion de rythme. L'accentuation correspond à la proéminence d'une syllabe par rapport à celles qui l'entourent. Cette proéminence est généralement due à une matérialité physique de par une augmentation (ou une variation contrastive) de un ou plusieurs paramètres acoustiques (durée, f0, intensité). Cela engendre alors une saillance de ces syllabes qui se démarquent des autres. Ces syllabes accentuées ne sont pas placées aléatoirement dans le flux de parole et sont soumises à différentes règles. Parmi elles, certaines sont présentes quelle que soit la langue parlée comme le fait d'éviter la présence de deux accents successifs (clashs accentuels). D'autres règles sont en revanche spécifiques à certaines langues ou groupes de langues. Par exemple, en français, l'accentuation est dite fixe car la syllabe finale d'un mot ou groupe de mot est systématiquement marquée d'un accent. En anglais, cette accentuation est dite libre car les accents ne marquent pas les mêmes syllabes d'un mot lexical à un autre. Ce "chef d'orchestre" qui expose les règles de l'accentuation est appelé *métrique* (Di Cristo et Hirst, 1996).

1.1.1 Structure métrique et prosodie des langues

Depuis des années, il a été montré qu'il est possible de distinguer différentes langues sans nécessairement reconnaître de mots, mais seulement en se basant sur les caractéristiques prosodiques et plus particulièrement rythmiques des langues. Nazzi et collab. (1998) par exemple ont montré que les nouveaux-nés étaient capables de discriminer plusieurs langues en se basant seulement sur des indices rythmiques (voir le chapitre 2 pour une description plus détaillées des études à ce sujet). En effet, la façon dont alternent les syllabes accentuées et inaccentuées est une caractéristique inhérente à la langue parlée. La place des syllabes accentuées dans le pied métrique (ensemble de syllabes constitué d'une syllabe accentuée et d'un nombre limité de syllabes inaccentuées, entre 1 et 3 en moyenne ; (Fant et collab., 1991)) correspond au gabarit métrique. Ainsi, en français, le pied métrique sera dit ïambique car son

accentuation primaire est situé à la fin (on parle de "tête métrique à droite"), tandis que l'accentuation primaire en anglais se situe au début et donc son pied est dit trochaïque ("tête métrique à gauche"). Bien que le français soit considéré comme ïambique, il n'est pas constitué uniquement d'un accent final primaire, il présente également une accentuation initiale potentielle qui marque le mot lexical au niveau sous-jacent et qui se situe au début des groupements prosodiques. Bien que d'un point de vue théorique, le français possède cette bipolarisation accentuelle, celle-ci ne se manifeste pas de manière systématique. La réalisation des accents est ainsi soumise à différentes contraintes. Par exemple l'accent initial est qualifié d'accent secondaire car il est possible qu'il disparaisse s'il se situe trop proche de l'accent final. Ainsi, dans le mot "Maison", l'accent initial sur "mai" cède sa place en surface au profit de l'accent primaire final "son". Ce genre de désaccentuation peut également se produire sur un accent primaire si deux accents finaux se suivent. Ce phénomène, appelé "clash accentuel" (Nespor et Vogel, 1989) provoque une réduction ou une disparition de l'un des accents. Dans le syntagme "la maison neuve", l'accent final de groupe sur le mot monosyllabique "neuve" implique une désaccentuation de la dernière syllabe de "maison" au profit d'une manifestation de l'accent initial. On se retrouve alors dans le cas d'un arc accentuel (Fónagy, 1980) où le premier et deuxième mots sont respectivement marqués par un accent initial et final.

Chaque langue possède donc des particularités rythmiques propres. L'étude des particularités métriques des différentes langues, combinée à l'observation de la tendance à la régularité des événements prosodiques (isochronie), a conduit pendant plusieurs décennies à la recherche de classification des langues en fonction du niveau sur lequel se porte cette régularité (syllabique vs. accentuelle). C'est tout d'abord Pike (1945) qui propose que les langues dont la durée inter accentuelle est régulière soient catégorisées comme langues à isochronie accentuelle. Tandis que les langues dont les durées des syllabes sont globalement similaires sont considérées comme langues à isochronie syllabique. L'anglais ferait alors partie de la première catégories et le français appartiendrait à la seconde. Cette classification a cependant été remise en question plusieurs fois au travers d'études acoustiques comme Dauer (1983) qui a observé que l'intervalle inter-accentuel moyen de l'anglais (isochronie accentuelle) était similaire à celui de l'espagnol (isochronie syllabique) ou encore par Astésano (2001); Fant et collab. (1991) qui ont montré que le français présentait également une certaine régularité inter-accentuelle. Wenk et Wioland (1982) quant à eux ont montré que les durées des syllabes en français n'étaient pas régulières mais étaient plutôt dépendantes de leur position dans le groupe prosodique et l'énoncé.

1.1.2 Le rôle de l'accentuation

L'accentuation découle des règles métriques des langues, puisqu'elle est la réalisation acoustique et la sensation perceptive du poids métrique. Elle est donc intimement liée au rythme linguistique. En fonction de la langue parlée, l'accentuation peut avoir

différentes fonctions plus ou moins importantes. En effet, selon Troubetzkoy et collab. (1939 ; cité par Di Cristo 2016b) l'accentuation a principalement 3 rôles :

— Un rôle culminatif qui indique que chaque mot lexical est marqué par au moins une syllabe accentuée.
— Une fonction distinctive dans certaines langues qui permet de différencier 2 mots partageant le même contenu segmental.
— Une fonction démarcative qui permet de localiser les frontières de début et fin de mots ou groupements de mots.

Ces fonctions sont plus ou moins importantes suivant la langue parlée. L'anglais qui est marqué par la présence d'un accent lexical utilise principalement les fonctions culminatives et distinctives tandis que le français est davantage marqué par la fonction démarcative étant donné que la place des accents est majoritairement fixe (comme expliqué dans la section 1.1.1).

Globalement, l'accentuation est intimement lié à la structuration du sens, au niveau lexical, syntaxique ou sémantico-pragmatique. Ainsi, bien que seules certaines langues impliquent un rôle de distinctivité lexicale à l'accentuation, il a été montré que le mauvais placement d'un accent dans un mot réduit significativement la vitesse de traitement du mot par les auditeurs (Cutler et collab., 1997), y compris dans une langue comme le français où l'accent n'est pas lié directement à la morphologie lexicale. Même dans une langue à accentuation fixe, un mauvais placement d'accent d'un point de vue métrique pourrait alors mener à un retard dans l'accès au sens du message. Cette assertion a été vérifiée pour le français dans les travaux de Astésano et collab. (2004) et Magne et collab. (2005). Ces études basées sur des signaux EEG ont montré qu'une syllabe métriquement forte placée sur la syllabe médiane d'un mot lexical donnait lieu à une négativité à 400 ms reflétant un ralentissement de l'accès au sens du mot, alors même que le mot est congruent sémantiquement. Cela suggère donc qu'un mauvais placement d'accent implique un retard dans le traitement et la compréhension des auditeurs.

1.1.3 L'accentuation comme structure hiérarchique

De nombreuses théories ont été établies afin de décrire les règles relatives à l'accentuation et plus généralement à la structure prosodique des langues. Concernant le français, et pour de multiples raisons que nous ne développerons pas ici (voir Astésano 2017, pour une discussion détaillée), la description prosodique a longtemps été dominée par les modèles intonosyntaxiques de Delattre (1966); Rossi et collab. (1981); Martin (1981) (cités par (Astésano, 2017)) qui se sont intéressés principalement à l'analyse de l'intonation dans son lien avec la structure syntaxique, sans référence claire à l'accentuation et la métrique. Une autre modélisation majeure de l'accentuation est l'approche de la métrique Auto Segmentale (AM) proposée par Pierrehumbert (1980) et adaptée au français par Jun et Fougeron (2000). Cette théorie basée sur la phonologie autosegmentale de Goldsmith (1976) étudie des segments tonals qui

décrivent les variations locales de f_0 au niveau syllabique et les mets en relation avec les contours intonatifs plus globaux indépendamment de la syntaxe.

Dans cette thèse, nous nous plaçons à la fois dans cette approche AM, pour les constituants prosodiques, et dans le cadre de la phonologie métrique qui est une théorie proposée par Liberman (1975) et Liberman et Prince (1977) et adaptée au français par Di Cristo (2000). L'approche métrique pose l'accentuation au coeur de la structuration prosodique en représentant les relations de hiérarchie accentuelle, indépendamment des corrélats acoustiques (rappelons que l'AM se fonde essentiellement sur les variations de f0 pour déterminer les types d'accents et de frontières). Cette théorie formalise l'alternance rythmique des syllabes accentuées ou non en ajoutant également une notion de poids métrique au centre de l'organisation accentuelle. La prosodie permet donc une structuration hiérarchique du flux de parole. Ainsi, ce flux continu peut être segmenté en unités de différentes tailles : les syllabes qui se regroupent elles-mêmes en mots puis en groupe de mots, en phrases et enfin en énoncés. Le nombre exact d'unités et leurs noms est un sujet encore aujourd'hui débattu (voir Shattuck-Hufnagel et Turk (1996) pour une revue des différents niveaux Di Cristo (2011) pour une vision comparée des niveaux prosodiques en fonction des approches théoriques). Nous considérons ici les niveaux suivants (du plus court au plus long) :

— La syllabe qui est le niveau où se matérialise l'accent de par des modulations physiques des sons.
— Le pied qui contient une syllabe accentuée suivie ou précédée (en fonction de la langue) d'un nombre fini de syllabes inaccentuées.
— Le mot prosodique (*pw*) qui est constitué d'un mot lexical (porteur de sens) et de son ou ses clitiques (ex : *un oiseau* ; niveau proposé pour le français par Astésano 2017)
— Le syntagme accentuel (AP) qui contient un mot prosodique et potentiellement son ou ses adjectifs (ex : *un petit oiseau bleu*). En l'absence d'adjectifs, il est possible que le niveau du *pw* et celui de l'AP représentent la même chose.
— Le syntagme intonatif (IP) est une unité contenant généralement plusieurs AP (Beckman, 1996) et qui est le domaine correspondant aux contours intonatifs (Delattre, 1966)
— L'énoncé qui correspond à minima à une 'phrase' complète et qui peut contenir une ou plusieurs IP.

En plus de ces différents niveaux de constituance, certaines études ont proposé un niveau intermédiaire situé entre l'AP et l'IP. Bien que son existence soit en partie acceptée dans certaines langues comme l'anglais (Beckman et Pierrehumbert, 1986) ou l'italien (D'Imperio, 2002), en français un tel niveau intermédiaire est largement sujet à controverse. En effet, même parmi les études qui soutiennent l'existence de ce niveau, sa structuration fait débat. Pour certains, ce syntagme intermédiaire (*ip, intermediate phrase*) serait essentiellement issue d'une certaine configuration syntaxique où une mise en exergue d'un syntagme est réalisée. À l'oral, les syntagmes nominaux font souvent l'objet d'un détachement à gauche avec reprise pronominale ; ce détachement créerait typiquement une frontière d'ip. Par exemple, dans la phrase *Voir un film au*

cinéma, ça l'intéresserait ?, le focus est effectué sur le groupe "Voir un film au cinéma" créant ainsi un détachement de cette partie qui résulte en la création d'une frontière d'ip après "cinéma".

Selon d'autres auteurs, l'*ip* est un niveau prosodique à part entière, qui présente une réalité physique de par un allongement syllabique et des mouvements de f_0 se situant à un niveau intermédiaire entre l'AP et l'IP (Michelas et D'Imperio, 2010b). Nous adoptons dans ce document le niveau de l'*ip* comme un véritable niveau prosodique intermédiaire qui dépend de la taille du constituant indépendemment de la syntaxe.

Un exemple des différents niveaux discutés précédemment peut être illustré par l'énoncé suivant :

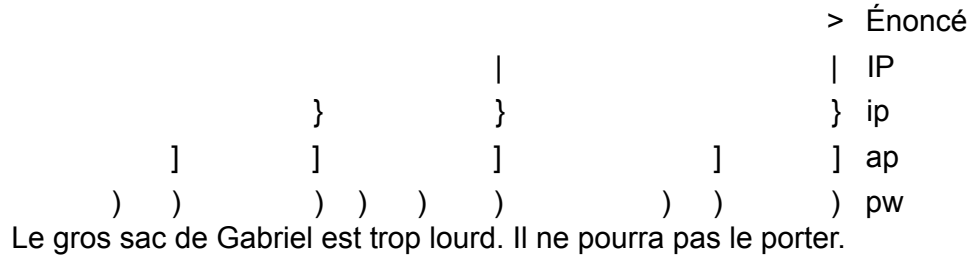

FIGURE 1.1 – Exemple des différents niveaux prosodiques que nous considérons : Mot prosodique (pw), syntagme accentuel (AP), syntagme intermédiaire (ip), syntagme intonatif (IP) et énoncé. Chaque niveau est illustré par un parenthésage à droite où tout ce qui précède la démarcation ( )] / } / | / >) jusqu'au signe précédent est inclus dans le niveau prosodique correspondant.

Bien qu'ici les différents niveaux sont clairement distinct les uns des autres, il est courant que certains de ces niveaux se confondent. Par exemple en supprimant l'adjectif "gros", le groupe "le sac" correspondrait alors à la fois au niveau du mot prosodique (*pw*) et au syntagme accentuel (ap). Cet exemple illustre les différents niveaux hiérarchiques qui organisent le flot parole. Cette organisation interagit donc avec le degré de proéminence des accents correspondant à chacun de ces niveaux. Le degré d'accentuation n'est donc pas lié uniquement au mot ou à la syllabe auquel il est rattaché, mais il dépend de la profondeur de la structure hiérarchique des constituants auxquels il appartient. Bien qu'il pourrait être tentant de penser que la structuration prosodique d'une phrase dépend principalement de sa syntaxe, on peut montrer qu'en réalité, les informations prosodiques et la hiérarchisation qu'elle génère, jouent un rôle déterminant, notamment dans les cas de désambiguïsation syntaxique. De façon plus concrète, prenons l'exemple suivant tiré de Aura (2012) :

"Les baguettes et les croissants chauds" (1)

Sans aucune autre information que la syntaxe, il est difficile de savoir si le mot "chauds" est rattaché uniquement aux "croissants" ou bien s'il inclue également "les baguettes". Le niveau de frontière prosodique peut alors rentrer en jeu et permettre à un auditeur de faciliter son accès au sens de la phrase. Ainsi, avec une structuration telle que

sur la partie haute de la figure 1.2, la frontière prosodique est plus forte après le mot "croissants" avec une accentuation plus marquée de la syllabe finale et potentiellement une pause indiquant la fin d'un niveau prosodique large (supérieur à l'AP), cela indiquant que l'adjectif concerne les deux noms précédents. Dans le cas où "chauds" est rattaché uniquement au mot "croissants", la structure prosodique devrait alors ressembler davantage à la partie basse de la figure 1.2.

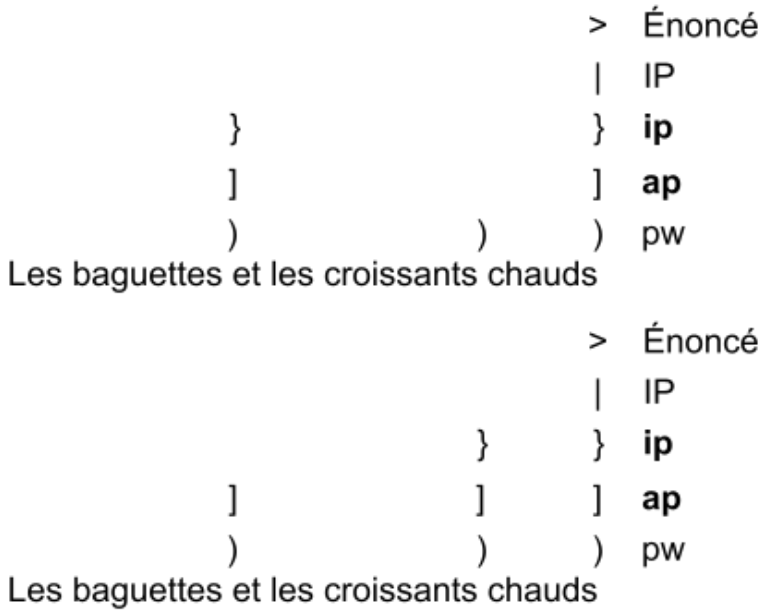

FIGURE 1.2 – Exemple de deux structurations prosodiques différentes pour la même phrase. En haut, l'adjectif "chauds" est rattaché seulement au mot "croissants" tandis qu'en dessous, "chauds" est rattaché à la fois aux mots "baguettes" et "croissants".

Comme indiqué dans la section 1.1.2, l'organisation hiérarchique de la parole montre que l'accent ne joue pas seulement un rôle au niveau lexical, mais également au niveau supra lexical qui permet d'améliorer l'accès au sens. Chaque accent a alors un niveau de proéminence plus ou moins élevé par rapport aux niveaux des autres accents. La mise en relation de l'accentuation à différents niveaux avec les niveaux prosodiques constitue le phrasé prosodique. En se basant sur ces propriétés, il est alors possible d'utiliser des modélisations formelles représentant le phrasé prosodique d'un énoncé avec ses niveaux d'accentuations et ses constituants.

1.1.4 Modélisations formelles de la théorie métrique

Afin de pouvoir rendre compte de la hiérarchisation de la parole dans le cadre de la théorie métrique, il a été proposé différentes modélisations afin d'avoir des représentations phonologiques des niveaux et des groupements prosodiques. La première modélisation formelle de la théorie métrique a été proposée par Liberman (1975) avec l'arbre métrique qui permet de représenter visuellement l'aspect hiérarchique de la parole en

montrant les degrés de proéminences relatifs aux différents niveaux de constituance prosodique. Cela est alors possible au travers d'un arbre binaire avec un étiquetage binaire fort/faible (s/w) des unités syllabiques et prosodiques. Cet étiquetage montre quelle unité possède un poids métrique plus élevé que l'autre. Par exemple, au niveau 2 de l'arbre métrique (groupes rythmiques/accentuels) dans la figure 1.3. et dans le cas de deux syntagmes, celui qui sera marqué par une frontière prosodique plus faible (disons ap) sera étiqueté w et le second (disons ip) sera étiqueté s. Un exemple d'arbre métrique est donné sur la figure 1.3.

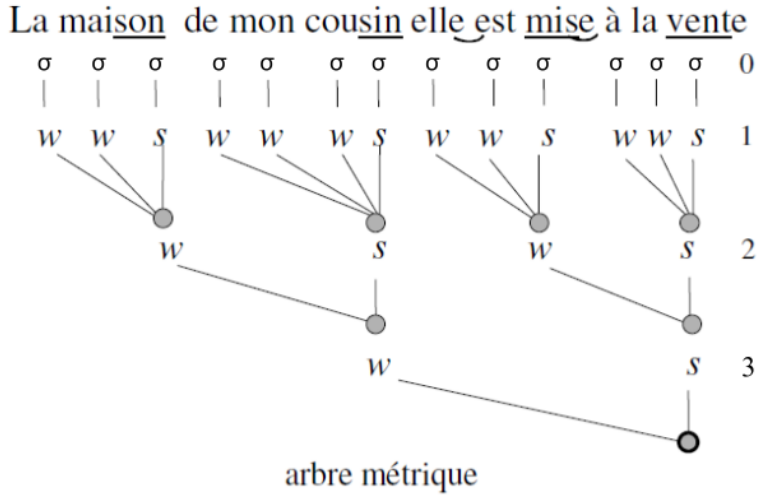

FIGURE 1.3 – Exemple d'arbre métrique tiré de Di Cristo (2016b, p.61). La phrase est initialement décomposée en syllabes σ (niveau 0), puis en syllabes accentuées s ou non w (niveau 1). Ces syllabes se regroupent pour former des niveaux prosodiques supérieurs eux-même faibles (w) ou forts (s) jusqu'à la racine de l'arbre qui englobe l'énoncé.

On peut voir ici que l'unité minimale est la syllabe avec une étiquette *s (strong)* pour les syllabes accentuées et *w (weak)* pour les syllabes inaccentuées. Plus on remonte dans l'arbre métrique vers sa racine, plus les unités prosodiques étudiées sont larges. Chacune de ces unités sont à leur tour étiquetées comme étant métriquement faibles ou fortes. Le choix pour déterminer quel segment est plus fort qu'un autre est généralement dépendant de la langue. Ici, le français est considéré comme étant une langue à accentuation fixe où les éléments dominants sont à la fin des niveaux prosodiques (nous détaillerons ces éléments dans la section 1.1.5). Dans cet exemple, les différents niveaux hiérarchiques représentés sont : les syllabes (niveau 1), le pied métrique (*pw* selon nous) au niveau 2, le syntagme accentuel (niveau 3) et enfin l'IP à la racine. Bien que l'arbre métrique soit utile pour rendre compte facilement de la hiérarchie accentuelle d'un énoncé, cette représentation présente néanmoins plusieurs limites. Premièrement, dans les niveaux supérieurs à la syllabe, la dimension temporelle de l'énoncé disparaît ce qui ne permet donc pas de retranscrire par exemple les règles d'alternances rythmiques des accents réalisés en surface comme les déplacements d'accents dans le cadre de collisions accentuelles (Selkirk, 1984; Di Cristo, 2004).

Une autre modélisation initialement proposée par Prince (1983) permet de corriger les défauts de l'arbre métrique en proposant une grille métrique (parenthésée ou non). Cette grille se base également sur un découpage en syllabes de l'énoncé où chacune de ces syllabes est marquée par un niveau de force d'accent représenté par des "∗". Une syllabes inaccentuée est donc représentée par une seule "∗" alors que les syllabes accentuées sont représentées par au moins deux "∗". Ce nombre augmente sur le plan vertical en fonction de la force du niveau de constituance associé. Le choix du niveau de force de la syllabe est défini comme suit : dans le cas de la grille métrique parenthésée, toutes les syllabes accentuées sont projetées au niveau supérieur de la grille ("∗∗"). Plus la syllabe en question marque la fin d'un constituant prosodique important, plus son niveau sera haut sur la grille. Un exemple de grille métrique est illustré sur la figure 1.4.

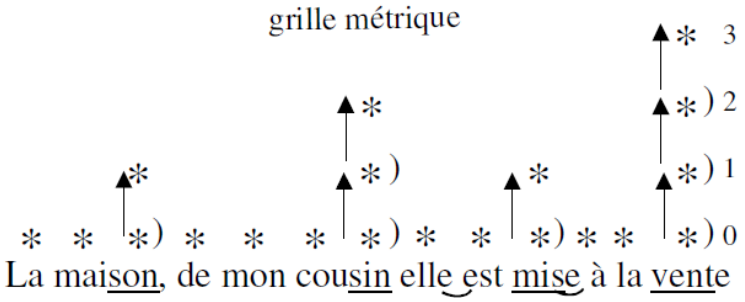

FIGURE 1.4 – Exemple de grille métrique parenthésée tirée de Di Cristo (2016b, p.61). En considérant les constituants définis en 1.1.3, on retrouve au niveau 0 les syllabes (marquées par des astérisques), les parenthèses indiquant les syllabes accentuées finales au niveau du pw ; ces syllabes accentuées se projettent au niveau 1 de l'ap ; le niveau 2 représente l'ip et enfin le 3 correspond à l'IP.

Comme illustré sur cette figure, la dernière syllabe qui marque la fin de l'énoncé possède le plus haut niveau d'accentuation et est donc marquée par quatre "∗". Le parenthésage proposé ici avec seulement les parenthèses fermantes signifie que toutes les syllabes précédentes (jusqu'à une autre parenthèse de même niveau) font partie du même groupe prosodique. Dans cet exemple, le mot "vente" est la syllabe ayant le parenthésage le plus haut qui se projette au niveau supérieur de la (macro) unité intonative : cela signifie donc que toute la phrase fait partie d'un même groupe qui est ici l'énoncé. De même, on retrouve deux unités de niveau intermédiaire qui selon Di Cristo sont des syntagmes intonatifs ou 'segments d'unités intonatives) : "La maison de mon cousin" et "elle est mise à la vente" mais qui correspondent selon notre vision davantage à des ip. Enfin, on retrouve quatre mots prosodiques *pw* "La maison", "de mon cousin", "elle est mise" et "à la vente", qui se confondent avec le niveau de l'ap (ou groupes accentuels chez Di Cristo).

Maintenant que nous avons vu comment la théorie métrique représente la structure de la parole de manière générale, il est temps de voir comment cette structuration s'applique en français en analysant le système accentuel de cette langue.

1.1.5 Le système accentuel français

L'accentuation est un phénomène hiérarchique où différents niveaux d'accentuation coexistent et sont en relation les uns avec les autres. Afin d'analyser la structure du français, il est indispensable de décrire ses particularités accentuelles.

Pendant longtemps, le français a été considéré comme étant une langue sans accent. Hjelmslev (1937) et Togeby (1965) (cité par Di Cristo (2016b, p. 12)) considèrent que le français n'est pas une langue à accentuation étant donné qu'il n'existe pas de distinction lexicale dépendante de l'accentuation. Hjelmslev affirme en revanche que des modulations existent bien en français mais il distingue ce phénomène de l'accent. Une étude importante qui a fait perdurer cette croyance de français sans accent est celle de Rossi (1980) qui s'est posé la question du domaine dans lequel s'applique l'accentuation en français. Il considère alors que le domaine prosodique de surface de l'accentuation n'est pas le mot lexical mais le groupe de mots. De fait, la matérialité de l'accentuation se confond avec celle du contour intonatif. Il parle alors de "syncrétisme" entre ces deux phénomènes. L'accent n'apportant selon Rossi pas plus d'information que les contours intonatifs marquant la fin de groupes prosodiques majeurs, il est alors naturel de considérer ces deux phénomènes comme une seule et même entité. Fónagy (1980) parle également de "chiasme acoustique" entre l'accent final et les mouvements d'intonation. Il considère que la mise en valeur des paramètres acoustiques de la syllabe accentuée se mélange avec l'intonation à la frontière des constituants.

Malgré la faiblesse des paramètres acoustiques de l'accent en français, certaines études ont pu montrer que les auditeurs francophones perçoivent malgré tout l'accent final. Smith (2009) a par exemple montré que les auditeurs percevaient bien les accents marquant la fin d'un groupe prosodique large tel que les IP. Astésano et collab. (2012) quant à eux ont observé que les auditeurs naïfs parvenaient à percevoir les proéminences finales, quel que soit le niveau de frontière prosodique. Ils concluent alors que la proéminence métrique finale en français est perçue indépendamment de la frontière prosodique marquée par un contour intonatif. Proéminence et intonation seraient donc bien deux phénomènes distincts en français.

Depuis longtemps l'accentuation du français est donc un sujet hautement débattu mais faire une revue détaillée de l'évolution des études à ce sujet serait hors du champ d'application de ce document. Pour une description historique de l'accentuation en français du XVI^{e} au XX^{e} siècle, une récente étude a été réalisée par Schweitzer et Dodane (2020). Astésano (2017) développe également l'évolution du statut de l'accentuation dans la phonologie du français en présentant en détail les différentes théories actuelles.

Nous nous plaçons ici dans le cadre majoritairement accepté où le français présente bien un système accentuel à part entière. Comme décrit dans la section 1.1.1, nous considérons notamment l'existence d'une accentuation finale de groupe en français comme étant un accent primaire (Fletcher, 1991; Di Cristo, 1999) et une accentuation initiale secondaire. Bien que la présence d'un accent final de groupe (FA) soit majoritairement admis, en revanche, la définition du niveau prosodique auquel se

rattache cet accent est beaucoup plus débattue. Le domaine de réalisation de l'accentuation ne serait pas le mot lexical, mais un groupe de plus haut niveau. Dans la majorité des modèles du français, ce groupe de mots correspondrait au niveau de l'AP (Jun et Fougeron, 2000). L'accent final correspondrait donc à la frontière droite du syntagme accentuel. Malgré tout, certaines études (Astésano, 2019) proposent de considérer davantage le niveau du mot prosodique (*pw*) comme niveau de réalisation de l'accentuation. En plus d'un accent final marqueur de frontière, un accent initial secondaire en début d'AP a été proposé. Cet accent marquerait le début des unités prosodiques sous certaines conditions, ainsi son aspect secondaire impliquerait qu'il puisse disparaître lors de la réalisation en surface au profit de l'accent final primaire (Pasdeloup, 1990; Di Cristo, 2016b). L'accent initial français est souvent vu comme un accent emphatique (d'insistance) qui ne se manifesterait que rarement. Il pourrait également jouer un rôle d'équilibrage rythmique afin de combler les vides accentuels de plusieurs syllabes (Rossi, 1980). Dans cette thèse, nous nous plaçons davantage dans la théorie selon laquelle l'accent initial est avant tout métrique et qu'il possède une force métrique similaire à celle de l'accent final (Astésano et Bertrand, 2016; Astésano, 2017). Il serait alors perçu avec la même force quel que soit le niveau de constituance auquel il est rattaché (Garnier et collab., 2016; Garnier, 2018) ce qui irait dans le sens d'un accent initial marqueur du niveau mot prosodique en surface (Astésano, 2019).

1.1.6 La matérialité de l'accent

Le français est donc marqué par une double accentuation initiale et finale présente à plusieurs niveaux hiérarchiques. Il est donc important de savoir comment ces différents types d'accents se manifestent d'un point de vue phonétique au travers des variations de fréquence fondamentale, de durée ou d'intensité.

Plusieurs travaux ont étudié les corrélats acoustiques liés aux différents types d'accents en français. L'accent final se manifesterait alors par des variations de f_0 dépendantes du domaine prosodique auquel il est rattaché (Di Cristo, 2016b). Dans son expérience, Post (2003) observe que l'ensemble des accents finaux rattachés au niveau de l'ap sont marqués par une montée de f_0. En revanche, les accents finaux d'énoncés (niveau de l'IP) seraient davantage marqués par une baisse de f_0 avec un ton de frontière bas. Une particularité intéressante des contours prosodiques du français que nous pouvons dénommer "inversion de pente" a été décrite par Delattre (1966) et Martin (1981). Ce phénomène concerne les mouvements de f_0 sur les syllabes finales d'ap successives au sein d'une IP. Nous pouvons alors observer des accents finaux porteurs de contours opposés d'une ap à la suivante pour instancier certaines relations syntaxiques internes à l'IP. Un ton bas sur l'accent final de la première ap impliquera alors un ton haut sur l'accent final de la seconde. Ce phénomène est illustré sur la figure 1.5.
Sur cet exemple, les deux ap "la soeur" et "de Jacques" présentent des courbes de f_0 opposées avec une courbe intonative opposée.

En plus de variations de f_0, l'accent final se manifeste en français également par un

FIGURE 1.5 – Illustration du phénomène d'inversion des pentes tirée de (Astésano, 2017, p.35)

allongement de durée de la syllabe proportionnel au niveau prosodique auquel il est rattaché (Parmenter et Blanc, 1933; Pasdeloup, 1990; Delais-Roussarie, 1995; Jun et Fougeron, 2000). Une syllabe rattachée à un accent final de syntagme intonatif (IP) serait alors jusqu'à deux fois plus longue qu'une syllabe non accentuée. La figure 1.6 illustre ce phénomène avec une étude (Jun et Fougeron, 2000) sur les durées moyennes des syllabes en fin de groupes prosodiques lors d'une lecture de la fable "La bise et le soleil".

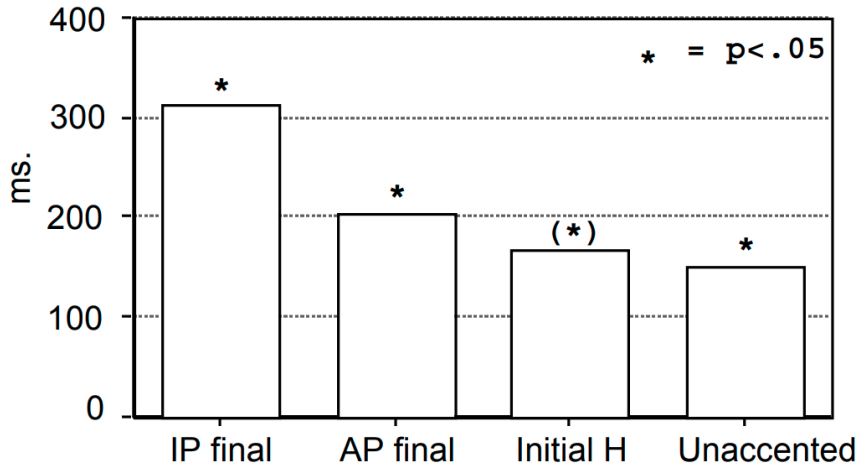

FIGURE 1.6 – Durée moyenne des syllabes associées aux accents finaux et initiaux dans différents contextes prosodiques. Figure extraite de (Jun et Fougeron, 2000)

Au-delà d'un allongement général de la syllabe accentuée, l'étude de (Astésano, 2001) s'est intéressée aux manifestations acoustiques des accents à un niveau infra-syllabique (décomposition en attaque, noyau + coda = rime). L'accent final serait alors marqué principalement par un allongement de sa rime proportionnel à la force de la frontière prosodique.

Concernant l'accent initial, il serait systématiquement marqué par un léger allongement de son attaque et une montée de f_0 asymétrique avec une hausse abrupte sur l'attaque et une baisse plus douce pouvant se répercuter sur les syllabes suivantes (Astésano, 2001, 2016). Certains auteurs notent que le pic de f_0 de AI n'est pas toujours aligné sur la première syllabe du mot ou du groupe. C'est ainsi que plusieurs auteurs parlent de l'accent initial comme d'un "marqueur flou de frontière" (Jun et Fougeron, 2000; Welby, 2003). Nous reviendrons sur ce terme et ses implications plus tard dans la section 1.2.3. Lorsque l'accent initial possède également un rôle emphatique (focus),

le pic de f_0 associé est plus élevé que pour un accent non-emphatique et une hausse de l'intensité est également observée (de Mareüil et collab., 2012). Au niveau de la durée de l'accent initial emphatique, une augmentation globale de la durée est observée, avec cependant un allongement de l'attaque plus important que celui de la rime (Astésano, 2001, 2016).

1.2 Le rythme comme manifestation de surface de la métrique

Dans la section précédente, nous avons essentiellement présenté d'un point de vue théorique comment s'organisent les accents dans la langue française au travers de la description de la métrique. Nous n'avons cependant pas abordé en détail la notion de rythme qui, bien qu'intimement liée aux notions de mètre et d'accentuation, a également une réalité psychologique et cognitive propre que l'on doit prendre en compte pour rendre compte de la programmation et la production de la parole

1.2.1 La métrique et le rythme de la parole : quelles différences ?

Le mètre et le rythme sont des notions qui ont parfois été confondues et dont les définitions ne sont pas concensuelles. La métrique peut ainsi correspondre à l´étude des régularités dans le cadre précis de la poésie versifiée (De Cornulier, 1995). Le rythme serait alors rattaché à la structure de la parole non versifiée. En se basant sur la théorie métrique de Di Cristo, on considère la métrique et le rythme comme deux niveaux cognitifs distincts : la métrique serait une représentation phonologique sous-jacente qui pose les règles de l'accentuation potentielle d'une langue en définissant la position des syllabes pouvant recevoir un accent par rapport aux syllabes inaccentuées. Le rythme quant à lui serait la manifestation en surface de ce dispositif sous-jacent mais avec une certaine flexibilité en s'adaptant aux différentes contraintes auxquelles le locuteur doit faire face comme la syntaxe, le contenu sémantique, le contexte pragmatique, ou encore le débit de parole. Une syllabe métriquement forte ne sera donc pas forcément pourvue d'un accent réalisé en surface. Cela peut être illustré par la notion d'eurythmie qui est un ensemble de règles qui sont appliquées en surface pour éviter le phénomène de "clash accentuels" décrit dans la section 1.1.1 ou encore de vide accentuel qui stipule qu'il n'est pas possible d'avoir un nombre trop grand de syllabes inaccentuées consécutives. Le débit de parole a également une incidence sur la réalisation des proéminences et la taille des groupes accentuels, cette dernière étant inversement proportionnelle au débit (Delais-Roussarie, 1995; Fougeron et Jun, 1998).

Ces processus permettent donc de mettre en évidence les différences entre la métrique sous-jacente et le rythme de surface. Il est néanmoins intéressant de noter que malgré cette variabilité de surface, Fant et collab. (1991) ont montré que quelle que soit la langue et son patron métrique (trochaïque, iambique), l'intervalle moyen entre deux

accents est généralement situé autour de 550 ms et est composé de 3 à 4 syllabes. Cette proposition a par la suite été observée en français par Astésano (2001).

1.2.2 La planification de la parole

Le rythme de la parole n'est pas parfaitement régulier, mais il peut se matérialiser à différent niveaux de la hiérarchie prosodique, et la régularité peut donc se trouver à un niveau ou à un autre. Notamment, Cummins et Port (1998a) proposent que le rythme de la parole est gouverné par un ensemble d'oscillateurs couplés qui seraient chacun responsable de la régularité d'un niveau prosodique particulier. Dans leur expérience Cummins et Port font répéter une phrase (en anglais) en s'alignant sur les battements d'un métronome de sorte à aligner la première et la dernière syllabe sur ces battements. Le métronome disparaît alors petit à petit, mais de façon surprenante, même lorsque les locuteurs n'étaient plus alignés temporellement avec le métronome, ils restaient en réalité dans un rapport harmonique de durée (par exemple 2 fois moins ou plus long que la durée initiale du métronome). Cette expérience suggère la présence de mécanismes (appelés oscillateurs) qui agissent conjointement lors de la production de parole, où chaque oscillateur serait responsable d'un niveau hiérarchique. Ainsi, il existerait un oscillateur responsable des syllabes qui serait inclus dans un autre responsable du niveau du pw (Arvaniti, 2021). Ces oscillateurs seraient alors tous ensembles synchronisés de sorte à ne pas avoir un décalage entre deux oscillateurs responsables de niveaux différents. Cette théorie pourrait alors expliquer les phénomènes de récurrence d'accents de façon régulière exposés dans la section précédente (1.2.1).

1.2.3 Perception du rythme

Le rythme n'est donc pas une application stricte de la structure métrique de l'énoncé. Le rythme a été défini par Astésano (2001) comme étant *l'organisation temporelle des proéminences*. Cette dénomination de "proéminences" englobe différentes notions comme l'aspect métrique, la réalisation en surface et également le côté perceptif. Certaines études se sont intéressées au lien entre les réalisations acoustiques et les niveaux de proéminence perçue. Un exemple concret pourrait être le cas de l'accent initial qui est considéré par certains comme un "marqueur flou de frontière" (Jun et Fougeron, 2000, 2002; Welby, 2003). Il est vrai que le pic de f_0 correspondant à un accent initial peut ne pas être parfaitement aligné avec la première syllabe porteuse d'accent, et parfois, il peut être placé sur la deuxième ou même troisième syllabe. Il est cependant intéressant de noter que l'étude de Astésano (2017) (également Astésano (2019)) a pu observer sur certains exemples que bien que le pic principal de f_0 soit aligné avec la deuxième syllabe, la syllabe perçue par des auditeurs naïfs comme étant la plus proéminente est bien la première et non la seconde. Ce phénomène est illustré sur la figure 1.7 avec la phrase "les bagatelles et les balivernes saugrenues". Le maximum de f_0 est situé sur les deuxièmes syllabes "ga" et "li" mais leurs scores de proéminence

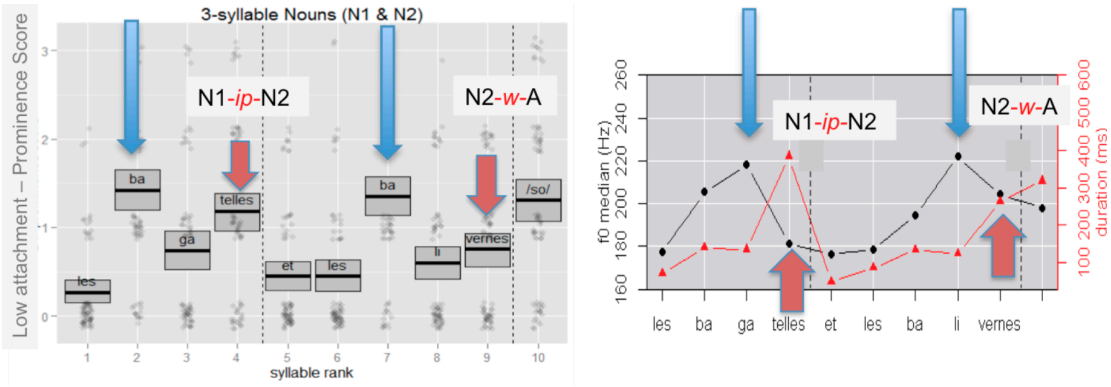

FIGURE 1.7 – Les scores de proéminences perçus sont indiqués à gauche, plus le score est haut, plus la syllabe en question est considérée comme proéminente. À droite, les courbes de durées des syllabes et de leur f_0 sont indiquées respectivement en rouge et en noir. Ici, nous voyons que le pic de f_0 sur "ga" de "bagatelles" (à droite) ne correspond pas à la perception de la proéminence sur la syllabe "ba" (à gauche). Figure extraite de (Astésano, 2017, p.104)

perçue sont relativement faibles tandis que les scores des premières syllabes sont les plus élevés. Ces résultats pourraient être expliqués par une perception liée à un "saut" de f_0 entre la syllabe inaccentuée liée à un mot fonction (par exemple un déterminant) et la première syllabe du mot lexical (saut perceptible si supérieur à 10%) ou bien par un processus top-down de reconnaissance du mot lexical par les auditeurs.

Dans le cadre du phénomène d'inversions de pentes décrit dans la section 1.1.6, Méndez et Astésano (2017) ont étudié la perception de la force de l'accent final. Selon cette étude, les accents finaux d'ap marqués par un ton bas sont détectés comme étant aussi proéminents que les accent finaux porteurs d'un ton haut. À des niveaux prosodiques plus élevés comme l'ip, ces accents finaux avec un ton bas pourraient même être perçus comme plus proéminents que les accents finaux usuels.

Les éléments métriquement forts ne sont en effet pas toujours réalisés en surface sur le plan phonétique mais peuvent parfois apparaître sous la forme de "mental beats" (Te Rietmolen, 2019) qui sont attendus par l'auditeur. C'est alors que même une pause dans la parole peut être considérée comme étant un beat fort (silent beat) et peut participer à la régularité du rythme de la parole étant donné que leurs durées sont généralement de rapport entier avec les constituants précédents ou suivants (Astésano, 2001). Le rythme ne peut pas seulement être considéré comme étant un phénomène linéaire issue de l'alternance faible/fort des syllabes d'un point de vue phonétique, il possède une forme propre et peut ainsi être considéré comme une Gestalt auditive (Fraisse, 1974), c'est à dire que c'est un ensemble de proéminences dissocié d'un fond d'évènements plus faibles. Le phénomène de Gestalt auditive se retrouve partout comme lorsque nous marchons où nous aurons tendance à regrouper perceptivement nos pas en un schéma de 3 ou 4 qui se répète ou encore en écoutant le bruit d'une horloge. Il est même possible de percevoir un rythme régulier là où il n'y en a pas. Il

est par exemple possible qu'un auditeur perçoive une séquence non-régulière d'évènements de plus en plus longs comme étant parfaitement régulière. Ce phénomène est appelé "time-shrinking" (Wagner, 2010).

Il est intéressant de noter que le tempo moteur spontané d'un individu pourrait jouer un rôle important dans la faculté à réaliser des groupements perceptifs. Ce tempo moteur se situe autour de 600 ms ce qui correspond également à l'intervalle inter accentuel moyen au niveau du pw ou de l'ap (Astésano, 2001; Astésano, 2022).

1.3 Conclusion du chapitre

Dans ce chapitre, nous avons vu que la prosodie est avant tout un élément qui permet de structurer la parole et de l'organiser tant du côté de l'auditeur que du locuteur. Dans un premier temps, nous nous sommes intéressé à la structure et aux différents rôles de l'accentuation. Nous avons vu que le français est composé d'une double accentuation initiale et finale dont le rôle est principalement démarcatif d'unité, mais également rythmique. Nous avons par la suite exposé la vision dans laquelle nous nous ancrons qui est celle de la phonologie métrique (Liberman, 1975; Liberman et Prince, 1977; Di Cristo, 2000) qui place l'accentuation au coeur de la structuration prosodique en proposant une formalisation de la hiérarchisation de l'accent. Ainsi, nous avons pu voir les différents niveaux prosodiques que nous considérons (pw, ap, ip, IP) ainsi que les outils formels de cette théorie : l'arbre et la grille métrique qui permettent d'avoir une représentation visuelle des relations inter-accentuelles. Également, nous avons pu voir que la manifestation acoustique de l'accentuation en français se traduisait principalement par des variations plus ou moins importantes de la durée des syllabes accentuées avec des structures infra-syllabiques différentes en fonction de l'accent et du constituant auquel il se rattache. Enfin, nous nous sommes intéressés au rythme qui peut être considéré comme la manifestation en surface d'un dispositif sous-jacent (la métrique). La manifestation du rythme se fait donc avec toutes les contraintes physiques, syntaxiques ou encore pragmatiques. Nous avons pu voir également que malgré tout, la recherche de la régularité du rythme pourrait être une réalité universelle avec une durée inter-accentuelle aux alentours de 550 ms dans de nombreuses langues (Fant et collab., 1991). Cette régularité pourrait être dû à une planification de la parole qui impose une régularité rythmique au travers d'oscillateurs couplés qui agissent à divers niveaux prosodiques (Cummins et Port, 1998b; Te Rietmolen, 2019). Notons cependant qu'accentuation et rythme n'ont pas forcément une réalité acoustique mesurable ; les auditeurs peuvent en effet percevoir une proéminence en l'absence de corrélats acoustiques associés généralement à cette proéminence ; de même, la régularité rythmique est davantage perceptive qu'acoustique (cf. 1.2.3). Trouver une modélisation automatique du rythme de la parole pertinente est donc d'autant plus difficile.

Maintenant que nous avons pu poser les fondements théoriques liés au rythme,

nous pouvons nous questionner quant à la faisabilité d'une modélisation automatique de ce dernier. Une bonne modélisation du rythme devra donc remplir plusieurs conditions : elle devra rendre compte de la régularité de la parole à plusieurs niveaux hiérarchiques, avoir une dimension temporelle et prendre en compte les variations de durée et/ou de f_0. La difficulté est donc de trouver la bonne méthode de recherche automatique des indices de l'accentuation et du rythme.

2

Les modélisations automatiques du rythme

Sommaire

Dans le chapitre précédent, nous avons vu que le rythme de la parole est un procédé complexe qui repose sur des fondements théoriques qui ne font pas toujours consensus au sein de la communauté. La création d'une modélisation de l'ensemble des spécificités du rythme est donc presque impossible pour le moment. Cependant, de nombreuses études ont été menées pour essayer d'extraire automatiquement des paramètres ou des modélisations qui représentent certains aspects du rythme langagier. Bien que, comme nous l'avons vu dans la section 1.2.3, le rythme ne peut être totalement modélisé au travers de l'acoustique du signal, nous allons ici nous intéresser aux études qui, à partir du signal acoustique, ont essayé d'extraire ou de modéliser des caractéristiques du rythme de manière (parfois quasi-)automatiques.

Nous allons dans un premier temps présenter des paramètres qui ont longtemps été utilisés dans un objectif de catégorisation automatique des langues, puis nous

regarderons quelques modélisations inspirées du domaine de la musique et qui ont été appliquées à l'étude de la parole. Enfin, nous nous intéresserons à des modélisations réalisées via l'analyse des modulations d'amplitude et de fréquence du signal, qui se rapprochent davantage des présupposés théoriques énoncés dans le chapitre précédent.

2.1 Les modèles du rythme pour l'identification des langues

L'identification des langues est un domaine de recherche encore aujourd'hui très prolixe. L'objectif est de détecter automatiquement quelle langue est parlée uniquement à partir de courts enregistrements de parole (identification des langues). Différentes méthodologies peuvent être employées en extrayant par exemple des paramètres acoustiques tels que les Coefficients Cepstraux en Fréquences Mel (*Mel-Frequency Cepstral Coefficients*, MFCC) (Davis et Mermelstein, 1980) et en utilisant des modélisations automatiques telles que les machines à vecteurs supports (*Support Vector Machines*, SVM) (Drucker et collab., 1997). L'identification de langues n'étant pas le sujet principal de cette thèse, nous nous renvoyons le lecteur vers la revue détaillée de ces méthodes faite par Ambikairajah et collab. (2011). Pour plus de détails sur les approches actuelles, nous pouvons citer les études de Lopez-Moreno et collab. (2014); Snyder et collab. (2018); Cai et collab. (2018, 2019) qui utilisent des modélisations à base de réseaux de neurones profonds. Dans ce chapitre, nous allons nous intéresser aux modélisations des langues n'utilisant que des informations sur le rythme de la parole.

2.1.1 Étude des durées vocaliques et consonantiques

Vers la fin des années 90, une nouvelle tendance dans l'identification des langues est apparue. Nazzi et collab. (1998) ont réalisé une étude portant sur l'identification des langues par des nouveaux-nés au travers d'enregistrements audio modifiés. Ces enregistrements ont été filtrés afin de ne conserver que les modulations lentes du signal. Ils ont alors démontré que les bébés étaient capables de distinguer des langues différentes telles que l'anglais et le japonnais uniquement à partir d'informations prosodiques. Ils n'étaient en revanche pas capables de faire la distinction entre le néerlandais et l'anglais par exemple. L'hypothèse avancée par les auteurs est que cette discrimination n'était possible qu'entre des langues de classes rythmiques différentes (dans le cadre de la catégorisation discutée en section 1.1.1). En effet, le japonais est considéré comme une langue à rythmicité moraïque (les durées entre deux mores sont équivalentes) alors que l'anglais et le néerlandais seraient des langues à rythmicité accentuelle (les durées entre deux syllabes accentuées sont équivalentes). À partir de cette étude, plusieurs auteurs se sont donc intéressés à l'analyse de la prosodie pour l'identification des langues.

Dans la continuité de l'étude de Nazzi et collab. (1998), Ramus et Mehler (1999) ont réalisé une expérience de perception dans laquelle ils ont extrait uniquement des informations rythmiques de diverses langues pour vérifier si des sujets français adultes étaient capables de distinguer des langues de différentes classes rythmiques. Afin de ne conserver que les aspects rythmiques (selon la vision d'une existence des classes d'isochronie) des enregistrements, ils ont resynthétisé les fichiers audio en remplaçant chaque voyelle par le son /a/ et chaque consonne par /s/. De plus, ils ont également forcé la f_0 à une valeur constante afin d'annuler les effets de l'intonation. Les auditeurs étaient alors généralement capables de discriminer des langues n'appartenant pas à la même classe d'isochronie. À partir de ces résultats, Ramus et collab. (1999) ont proposé un ensemble de paramètres par phrase en se basant sur les durées des intervalles vocaliques et consonantiques. Les intervalles vocaliques sont définis comme étant l'intervalle entre l'attaque et la fin d'une succession de voyelles (ou d'une seule voyelle si elle est isolée par des pauses ou des consonnes). De même un intervalle consonantique représente la durée d'une consonne ou d'un ensemble de consonnes successives. Voici le détail des paramètres extraits sur chaque phrase :

— $\%V$ qui est la proportion d'intervalles vocaliques dans la phrase (durée des intervalles divisée par la durée de la phrase)
— $\%C$ est la proportion d'intervalles consonantiques, mais n'est généralement pas utilisé car il est réciproque de $\%V$
— ΔV l'écart type des durées des intervalles vocaliques
— ΔC l'écart type des durées des intervalles consonantiques (qui cette fois apporte une information complémentaire à ΔV).

Ces paramètres ont alors été extraits sur un ensemble de plusieurs langues afin de vérifier s'il était possible de rendre compte des différentes catégories d'isochronie. Le tableau 2.1 décrit les résultats obtenus. On peut alors remarquer que la proportion

TABLE 2.1 – Exemple de paramètres rythmiques extraits sur un ensemble de langues de différentes classes rythmiques. $(\%V)$, est la proportion des durées vocaliques, (ΔV) l'écart type des durées des intervalles vocaliques, ΔC l'écart type des durées des intervalles consonantiques. Les écarts-types sont indiqués entre parenthèses. Tableau adapté de (Ramus et collab., 1999)

Langue	Classe rythmique	$\%V$	$\Delta V(\times 100)$	$\Delta C(\times 100)$
Anglais (EN)	Accentuelle	40.1(5.4)	4.64(1.25)	5.35(1.63)
Polonais (PO)	Accentuelle	41.0(3.4)	2.51(0.67)	5.14(1.18)
Néerlandais (DU)	Accentuelle	42.3(4.2)	4.23(0.93)	5.33(1.5)
Français (FR)	Syllabique	43.6(4.5)	3.78(1.21)	4.39(0.74)
Espagnol (SP)	Syllabique	43.8(4.0)	3.32(1.0)	4.74(0.85)
Italien (IT)	Syllabiques	45.2(3.9)	4.00(1.05)	4.81(0.89)
Catalan (CA)	Syllabiques	45.6(5.4)	3.68(1.44)	4.52(0.86)
Japonais (JA)	Moraïque	53.1(3.4)	4.02(0.58)	3.56(0.74)

d'intervalles vocaliques ($\%V$) semble plus importante dans les langues à rythmicité syllabique que dans celles à rythmicité accentuelles. Ce résultat pourrait alors s'expliquer par le phénomène de réduction vocalique qui dans les langues comme l'anglais

implique que les voyelles des syllabes inaccentuées ont tendance à se rapprocher de la voyelle neutre (qui correspond au /ə/ du mot "le" en français) ce qui peut parfois engendrer une réduction de leur durée (Fourakis, 1991). De plus, l'écart type des durées inter-consonantique (ΔC) semble plus court pour les langues à rythmicité syllabique ce qui pourrait être un argument en faveur de cette classification. En utilisant ces deux paramètres les plus discriminants, il est alors possible de représenter les langues dans un espace à deux dimensions comme représenté sur la figure 2.1. Sur cette image, il est alors possible de distinguer les trois catégories de langues avec les langues à rythmicité accentuelle en haut à gauche, les langues à rythmicité syllabique au milieu et le japonais qui représente les langues à rythmicité moraïques en bas à droite.

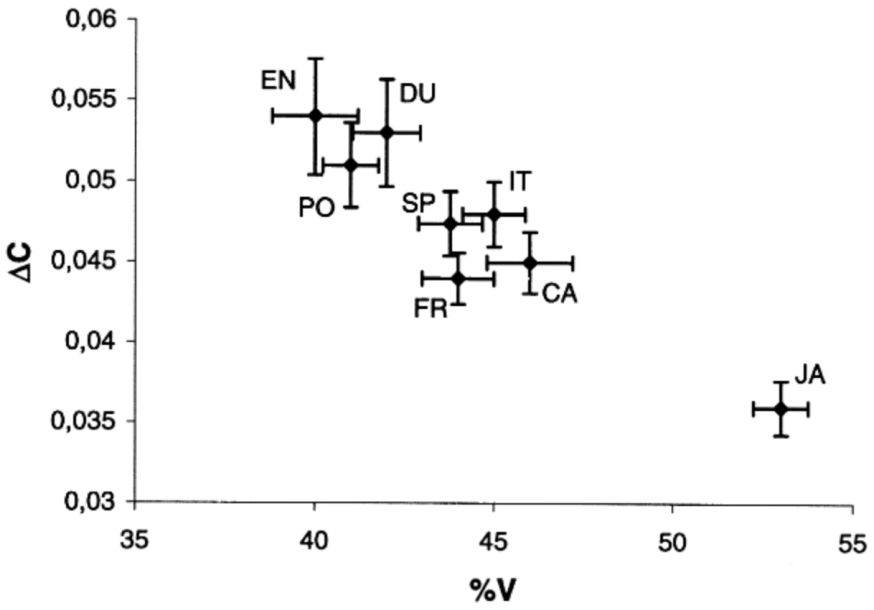

FIGURE 2.1 – Représentation des mesures des paramètres $\%V$ et ΔC pour chacune des langues étudiées sur un corpus de parole lue. Les barres d'erreurs correspondent à $\pm$ l'écart-type des paramètres. Figure tirée de (Ramus et collab., 1999, p.273).

Les indices rythmiques proposés par Ramus et collab. (1999) sont donc intéressants pour caractériser en partie le rythme de diverses langues. Cependant, bien que ces mesures semblent simples à calculer, elles nécessitent d'avoir à disposition une transcription phonétique (ou à minima des segments vocaliques et consonantiques), ce qui n'était pas facilement disponible à l'époque de la publication de ces travaux. Un intérêt pour l'automatisation de l'extraction d'une segmentation fiable du signal s'est alors développé avec notamment les travaux de Pellegrino (1998) proposant une modélisation multilingue des segments vocaliques, ce qui a permis de proposer une modélisation de pseudo-syllabes, et de modéliser le rythme à partir de certains des paramètres de Ramus, de manière automatique (Pellegrino et collab., 2002; Farinas, 2002; Farinas et collab., 2005; Rouas et collab., 2005).

Bien que ces paramètres semblent pertinents pour catégoriser différentes langues, ils présentent un problème majeur. Ces indices bruts sont pertinents dans un cadre

de parole contrôlée comme de la parole lue, mais dans le cas de parole plus spontanée où il existe des hésitations et des variations du débit de parole, les durées vocaliques et consonantiques sont alors soumises à une grande variabilité intra et inter locuteurs.

2.1.2 Normalisation par le débit de parole

Selon Dellwo et Wagner (2015) et Dellwo (2006), les variations de débit de parole influencent fortement les paramètres de Ramus et collab. (1999) et plus particulièrement ΔC qui est fortement influencé par ces variations que ce soit chez un même individu auquel il est demandé de modifier son débit de parole ou bien chez deux individus de débit différents. Pellegrino et collab. (2004) ont également pu observer que dans un cadre de parole spontanée, des différences significatives pouvaient apparaître entre différentes langues. Il est donc primordial de corriger ces paramètres afin qu'ils ne tiennent pas compte des variations de vitesse d'élocution. Ling et collab. (2000) ont par exemple proposé un nouvel indice appelé *Pairwise Variability Index* (PVI) qui se base sur les durées relatives de segments vocaliques et inter vocaliques successifs. Une version brute (rPVI) et une version normalisée (nPVI) par rapport au débit ont été établies. Le paramètre brut rPVI est calculé comme indiqué dans l'équation 2.1 :

$$rPVI = 100 \times \sum_{k=1}^{m-1} |d_k - d_{k+1}| \tag{2.1}$$

Où m est le nombre d'intervalles vocaliques ou intervocaliques et d_k est la durée du k-ième intervalle. Pour obtenir une version normalisée, il est alors nécessaire de diviser ces valeurs de durées par une durée moyenne comme décrit dans l'équation 2.2 :

$$nPVI = 100 \times \left[\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| /(m-1) \right] \tag{2.2}$$

Dans la même optique, Dellwo (2006) ont proposé une nouvelle mesure adaptée du ΔC afin de limiter les effets de variations du débit. Ce nouveau paramètre appelé $varco\Delta C$ est défini comme :

$$varco\Delta C = \frac{\Delta C \times 100}{meanC} \tag{2.3}$$

Où $meanC$ est la durée moyenne des intervalles consonantiques.

À partir de ces nouveaux indices, Grabe et Low (2002) ont alors comparé les PVI avec les paramètres proposés par Ramus et collab. (1999) décrits dans la partie précédente (2.1.1) afin de vérifier qu'ils permettent bien de discriminer les catégories de langues de la même façon. Les résultats obtenus sont illustrés dans la figure 2.2 où nous retrouvons bien les langues à rythmicité accentuelles, les langues à isochronie syllabiques en bas à gauche et le japonais qui est une langue à isochronie moraïque en bas à droite.

Concernant le paramètre $varco\Delta C$, l'étude de Dellwo (2006) a montré qu'en faisant

varier les productions des locuteurs (en terme de vitesse et de f_0, une grande variabilité était observée pour le paramètre de ΔC. Cette variabilité était alors atténuée en utilisant $varco\Delta C$. En effet, la distinction entre l'anglais et l'allemand (toutes deux à isochronie accentuelle) était impossible avec ΔC à cause de la variabilité intra locuteur, tandis qu'avec $varco\Delta C$, la distinction entre ces deux langues était davantage marquée. Il est également important de souligner que les valeurs de ΔC entre anglais et français étaient parfois confondues en fonction du type de parole, mais avec $varco\Delta C$, la séparation entre ces deux langues était alors plus significative.

Il est tout de même important de noter que dans la majorité des études citées dans le paragraphe précédent, les langues étudiées sont des langues dîtes prototypiques de chacune des catégories d'isochronie, mais il existe également de nombreuses langues qui présentent des caractéristiques mixtes qui n'appartiennent pas à une catégorie précise. C'est le cas par exemple du Catalan ou du Polonais qui n'appartiennent à aucune de ces classes. Lorsque les paramètres définis précédemment sont utilisés pour séparer ces langues, la classification devient alors plus complexe et nous observons que plusieurs langues se retrouvent entre les différentes classes (Grabe et Low, 2002). Ces observations pourraient alors appuyer la proposition de Dauer (1983) qui suggère que les classes de rythmes n'existent pas concrètement, mais qu'il s'agit davantage d'un continuum qui défini les langues comme étant plus ou moins isochrone au niveau syllabique, accentuel ou moraïque.

En plus du domaine de l'identification des langues, les indices du rythme que nous avons détaillés dans la section 2.1 ont été utilisés dans le cadre de la distinction des styles de paroles. Ce domaine de recherche consiste à classer les différents types de

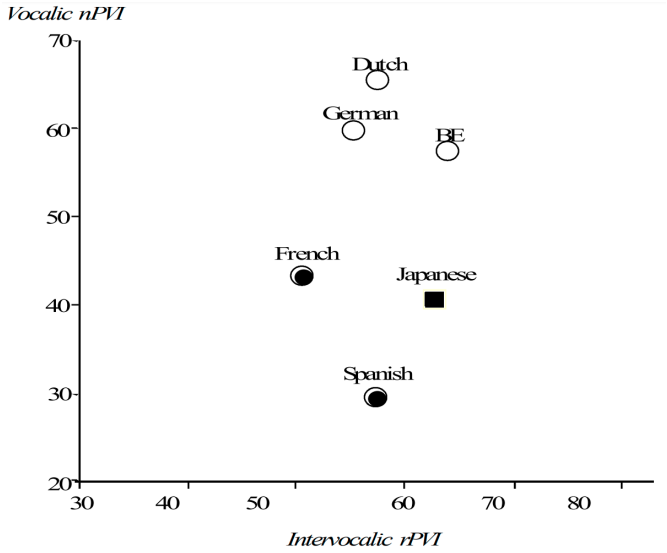

FIGURE 2.2 – Représentation des mesures des paramètres $nPVI$ et $rPVI$ pour chacune des langues étudiées sur un corpus de parole lue. Les ronds blancs sont les langues à isochronie accentuelle, les rond noirs représentent les langues à isochronie syllabique, et le japonais (moraïque) est représentée par un carré. Figure tirée de (Grabe et Low, 2002, p.6).

parole existants comme la parole lue, chantée, spontanée, ou encore dans des discours télévisés. Certains auteurs se sont alors intéressés à l'utilisation des indices du rythme pour détecter automatiquement ces styles de paroles. Généralement, des informations comme le débit de parole ou le débit articulatoire (nombre de syllabes / durée de parole sans pauses) sont utilisées (Simon et collab., 2010; Goldman et collab., 2009). Certains auteurs ont néanmoins également utilisé les indices tels que les paramètres $varco$, ΔC, $\%V$ (Prsir et collab., 2014). Ils ont ainsi pu montrer par exemple que le $\%V$ était plus élevé dans les discours religieux tandis qu'il semble être plus faible pour les prévisions météo ou la lecture de textes.

2.2 Les modèles du rythme en musique

Bien que les paramètres évoqués précédemment permettent une différenciation relativement satisfaisante des langues, ils ne permettent pas de rendre compte de la complexité structurelle du rythme car ils ne s'intéressent qu'à une analyse bas niveau du signal via des mesures de durées segmentales. Ces paramètres ne prennent pas non plus en compte la structuration rythmique interne des langues. Or, il nous intéresse de pouvoir comparer les variations de structuration rythmique dans une même langue, chez un même locuteur ou un pool de locuteurs. Certaines modélisations issues de la recherche musicale permettent de prendre en compte la régularité rythmique à différents niveaux de la structure musicale. Il pourrait être intéressant de les appliquer à la parole. Elles présentent l'avantage de rendre compte de l'aspect hiérarchique de la métrique et du rythme.

2.2.1 Extraction du tempo dans la musique

Le rythme au sens général dans la musique est un concept très vaste, il est donc très difficile de proposer une modélisation qui rende compte des différents aspects métriques, acoustiques ou hiérarchiques. Un élément plus simple à étudier est l'analyse des battements (*beats*) qui correspondent à des impulsions apparaissant de façon régulière et dont la fréquence correspond au *tempo* (Scheirer, 1998). Ce *tempo* peut apparaître à différents niveaux métriques (tout comme en parole, voir 1.1.3), et il existe généralement un niveau métrique dominant dans lequel le tempo correspond à la fréquence à laquelle un humain frapperait dans ses mains en écoutant la musique.

L'extraction automatique du tempo a alors été un sujet vastement exploré. Dans un premier temps, l'extraction du tempo se faisait en analysant les régularités dans les débuts de notes (*onsets*). Par exemple, dans les études de Desain (1992) et Large et Kolen (1994), les *onsets* étaient préalablement annotés. Desain (1992) traite séquentiellement les onsets en entrées et compare les durées entre les onsets afin de voir si certaines durées apparaissent plus souvent que d'autres. Large et Kolen (1994) ont quant à eux essayé de faire correspondre une combinaison d'oscillateurs non-linéaires au dessus des onsets en entrée. Concrètement, un oscillateur est un système

qui produit un signal d'amplitude (en électronique, cette amplitude serait un courant électrique) périodique avec une fréquence qui peut être fixe ou variable. L'idée était donc de modéliser la succession d'onsets comme étant le résultat d'une combinaison d'oscillateurs de fréquences différentes. Cette idée correspondrait notamment aux théories avancées dans la section 1.2.2.

Bien que ces modélisations soient pertinentes, elles se basent sur une annotation souvent manuelle des onsets. Afin de remédier à cela, des méthodes de détection automatique de ces derniers ont été développées. Ainsi, Dixon (2001) a par exemple réalisé des estimations des positions d'onsets en se basant sur les pics d'énergie du signal. Dans la même idée, Miguel Alonso et Richard (2004) recherchent les onsets possibles en utilisant des différences d'énergies dans diverses bandes de fréquences du signal. Les potentiels onsets modélisés par des pics aux durées sélectionnées sont alors étudiés en analysant les durées entre les pics. Cette analyse des durées peut alors être réalisée de plusieurs façons.

L'une des possibilité est par exemple de s'intéresser aux différentes fréquences d'apparition des pics (onsets). Pour cela, il est possible d'utiliser une fonction de mesure de la périodicité d'un signal, comme par exemple la transformée de Fourier discrète (Pampalk et collab., 2002) qui permet de décomposer un signal en une combinaison de sinusoïdes de différentes fréquences (Bracewell et Bracewell, 1986) ou encore l'autocorrélation (Brown, 1993; Scheirer et Slaney, 1997). Cette dernière méthode mesure la similarité d'un signal avec une version décalée de celui-ci. En faisant varier ce décalage τ, nous pouvons alors mesurer la similitude entre le signal $S(t)$ et sa version décalée $S(t+\tau$. Les décalages fournissant un signal proche de l'original nous indiquent ainsi le tempo (et ses multiples). Nous pouvons en déduire la fréquences d'apparition des onsets des notes. Il est ensuite possible d'extraire les différents candidats potentiels pour estimer le tempo en observant les pics de fréquences les plus élevés. Un schéma de l'architecture générale d'extraction du tempo est proposé en figure 2.3.

Comme nous pouvons l'observer sur la figure, il peut cependant être difficile de trouver la véritable valeur du tempo car l'analyse fréquentielle peut amener à montrer des pics à plusieurs fréquences. On peut alors être amenés à utiliser différentes stratégies afin de sélectionner le meilleur candidat possible (voir Gouyon et Dixon (2005) pour une revue des méthodes d'extraction). Une des techniques les plus utilisées consiste à utiliser un "peigne spectral" (Le Coz, 2014) qui consiste pour plusieurs fréquences τ potentielles à calculer :

$$T = argmax_\tau \left[S\left(\frac{\tau}{3}\right) + S\left(\frac{\tau}{2}\right) + S\left(\tau\right) + S\left(2\tau\right) + S\left(3\tau\right) \right]$$

Où S est la fonction de périodicité choisie et T le tempo final retenu. Cette méthode permet donc de sélectionner la fréquence qui génère le plus d'énergie avec ses harmoniques et ainsi éviter d'estimer un tempo deux fois trop grand ou deux fois trop petit.

Ces méthodes permettent donc d'estimer efficacement le tempo d'une musique.

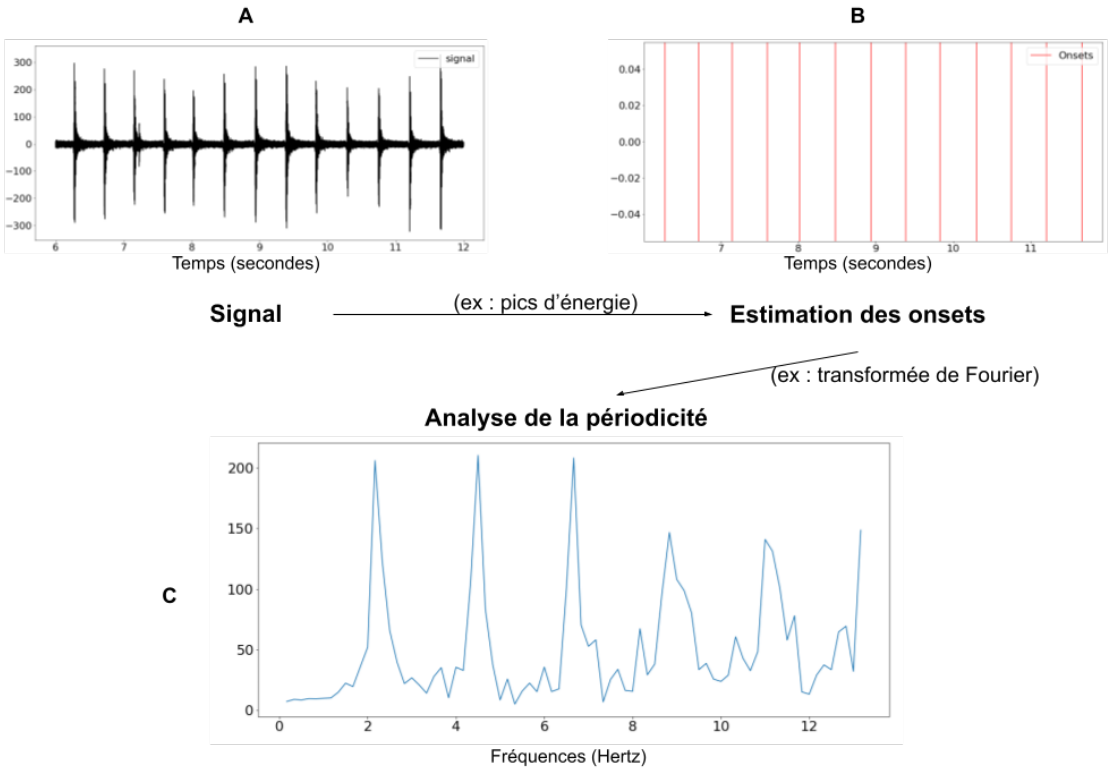

FIGURE 2.3 – Illustration d'une méthodologie couramment employée pour la détection automatique du tempo en musique. A : le signal initial, ici un tapping régulier avec environ 2 tapes par secondes (2 hertz). B : les durées à laquelle les onsets ont été estimés. C : la transformée de Fourier des pics correspondants aux onsets. Ici, la fréquence du premier pic de la transformée de Fourier correspond au tempo.

Cependant, dans un cadre comme de la parole ou plus simplement pour une musique dont le tempo évolue au cours du temps, il n'est pas possible de les utiliser.

2.2.2 Le tempogramme

Afin de prendre en compte le côté dynamique du rythme (évolution dans le temps) ainsi que sa hiérarchisation, il est nécessaire d'utiliser des modélisations plus adaptées. Pour cela, un outil initialement développé pour l'étude de la musique peut être utilisé afin de modéliser le rythme de la parole : le tempogramme. Il existe deux versions du tempogramme publiées la même année qui sont fortement similaires : celui de Le Coz et collab. (2010); Le Coz (2014) et le tempogramme cyclique proposé par Grosche et collab. (2010). Ces deux versions diffèrent principalement dans la première partie de leur implémentation qui consiste à choisir la détection d'accentuation dans le signal (caractérisée par les onsets dans la section 2.2.1 précédente). Le principe général du tempogramme est d'étudier les régularités d'apparition d'un évènement particulier au cours du temps. Initialement, le tempogramme est un algorithme utilisé

afin de détecter et suivre le tempo de musiques. Le principe du tempogramme de Le Coz est le suivant :

— Segmenter le signal brut en marquant des frontières correspondant à des évènements particuliers (nous y reviendrons)
— Pondérer les frontières obtenues en fonction de l'énergie spectrale avant et après celles-ci
— Utiliser une transformée de Fourier (TFD) sur ces pics (frontières)
— Répéter les étapes précédentes sur des fenêtres glissantes de quelques secondes (environ 3 secondes)

La première étape consiste donc à réaliser une segmentation du signal. La segmentation utilisée par Le Coz et collab. (2010) est la divergence *forward-backward* qui a été créée par Andre-Obrecht (1988) et qui permet de segmenter le signal en séquences d'éléments stables. Dans le cadre de la musique, les éléments stables peuvent être les quatre étapes d'une note de musique du point de vue de l'enveloppe du signal : l'attaque qui correspond au temps que le son va mettre pour atteindre son amplitude maximale, le déclin qui décrit le temps où le son redescend de son niveau maximal vers un niveau plus stable appelé le maintien et enfin la chute qui correspond au temps que la note met à atteindre une amplitude nulle après avoir relâché la note. La segmentation *forward-backward* effectue la segmentation en élément stable au travers d'un modèle autorégressif, c'est à dire un modèle qui cherche à modéliser les données futures d'un signal à partir d'une combinaison linéaire de ses éléments précédents. Ainsi, s'il est possible de prédire la suite d'un segment de signal à partir de ses éléments précédents, on peut alors considérer que le segment est stable. En revanche, si la prédiction est trop éloignée de la réalité, alors une frontière est créée. Cela se traduit formellement par le modèle autorégressif suivant :

$$\begin{cases} y_n = \sum_{i=1}^{k} a_i y_{n-i} + e_n \\ var(e_n) = \sigma_n^2 \end{cases} \tag{2.4}$$

Où y_n est le n-ième élément du signal, k est l'ordre du modèle, a_i les paramètres du modèle et σ un bruit gaussien. Ce modèle est calculé pour détecter les ruptures dans le signal et permet ainsi d'obtenir une segmentation comme illustrée sur la figure 2.4

Une fois la segmentation obtenue, une pondération des pics (frontières) est réalisée en calculant la différence entre l'énergie du spectre 20 ms après et 20 ms avant chaque frontière. Ces pics ainsi calculés forment alors le matériau de base pour analyser plus tard la périodicité du signal. Dans l'implémentation de Grosche et collab. (2010), ce matériau est calculé différemment. Dans leur étude, ils effectuent une Transformée de Fourier sur des courtes fenêtres glissantes de 20 ms et calculent la dérivée de l'énergie de ces spectres. Seules les dérivées positives sont conservées afin de ne garder que les augmentations d'énergies qui pourraient correspondre à l'onset d'une note.

À partir de ces diracs (pics) (approche Le Coz) ou de cette courbe (approche Grosche), une transformée de Fourier est utilisée dans le but de détecter des régularités dans les apparitions des pics. Un exemple de Transformée de Fourier sur la segmentation d'une

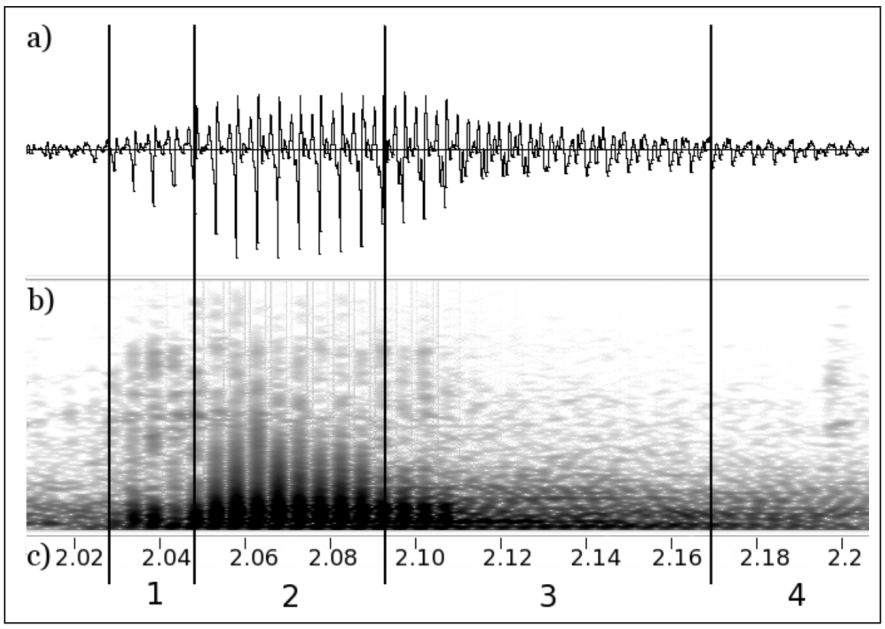

FIGURE 2.4 – Exemple de la segmentation forward-backward sur une note de musique d'un trombone. Les segments 1, 2 et 3 correspondent respectivement aux phases d'attaques, de maintien et de chute. Figure extraite de Le Coz et collab. (2010, p.28).

musique complète est donné dans la figure 2.5. Dans cette figure, le spectre génère plusieurs pics situés à des fréquences multiples l'une de l'autre. Ainsi, il apparaît différents pics correspondant au tempo principal recherché, mais également au double et triple du tempo. La détection du tempo principal peut alors se faire en utilisant un peigne spectral comme décrit dans la section précédente (2.2.1).

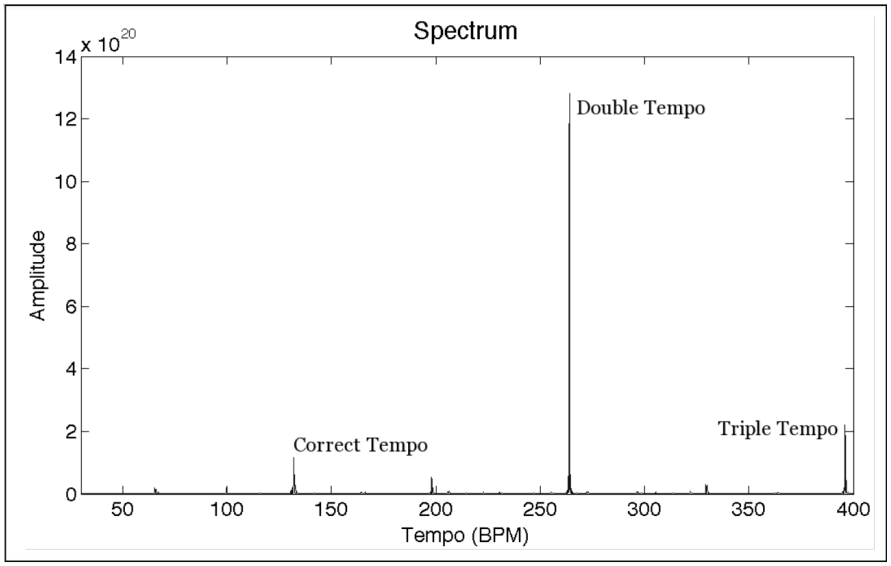

FIGURE 2.5 – Exemple de Transformée de Fourier d'une segmentation forward-backward sur une musique entière. L'abscisse correspond aux fréquences en battements par minutes (BPM). Figure extraite de Le Coz et collab. (2010, p.29)

Cette méthode permet donc de calculer assez précisément le tempo d'une musique. En revanche, si jamais le tempo évolue au cours de la musique, en réalisant les calculs sur l'ensemble du morceau, il est difficile de trouver quels sont les différents tempos mis en jeux. Une solution possible est alors de passer de la représentation spectrale à une représentation temps-fréquences au travers d'un calcul du spectre sur une fenêtre glissante du signal. Cela résulte en un spectrogramme du rythme appelé le tempogramme. Une illustration de tempogramme calculé sur une fenêtre glissante de 3 secondes est donné sur la figure 2.6 où l'évolution du tempo principal dans le signal est mis en évidence par l'apparition d'une bande rouge aux alentours de 1 Hertz (en ordonnées) pendant les 40 premières secondes signal, puis par une rupture soudaine qui laisse apparaître un nouveau tempo à 1,6 Hz à partir de 45 secondes.

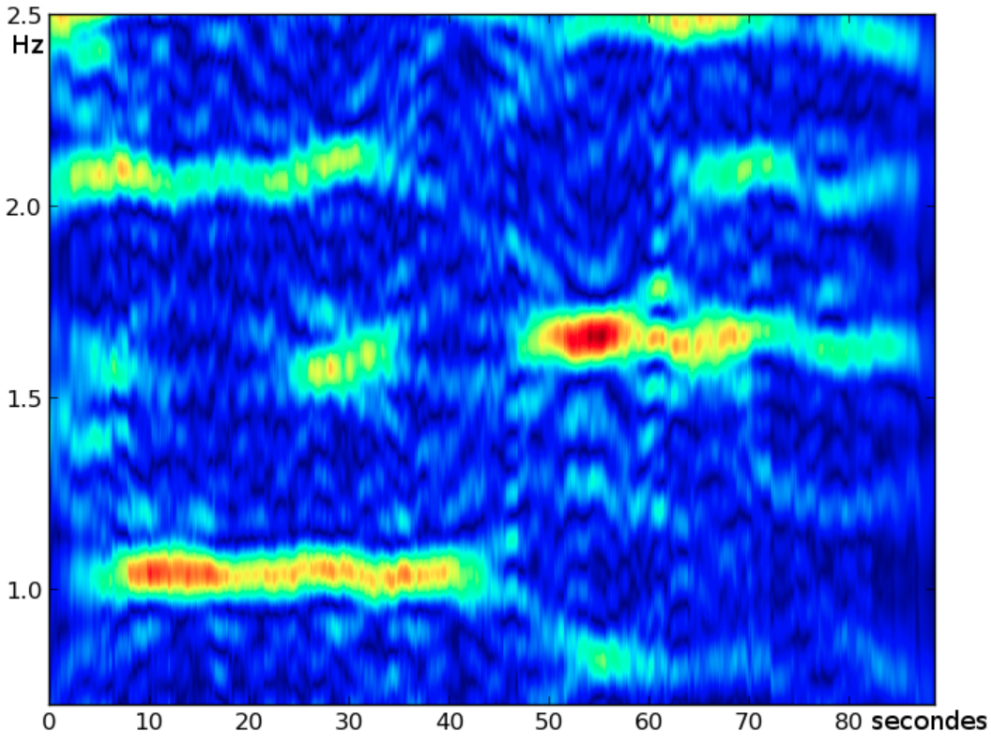

FIGURE 2.6 – Exemple de tempogramme sur un signal dont le rythme principal est à environ 1 Hz durant les 45 premières secondes puis à 1,6 Hz. Les fréquences (en BPM) sont indiquées en ordonnées en fonction du temps (en secondes). Plus une zone tend vers le rouge, plus l'énergie dans cette bande de fréquences est importante. Figure extraite de `https://www.irit.fr/SAMOVA/site/research/analysis/rhythm-estimation/`.

Cette modélisation du tempo musical pourrait donc être facilement adaptée à l'étude du signal de parole en sélectionnant une segmentation différente basée par exemple sur les segments vocaliques et inter-vocaliques comme développé dans la section 2.1. Cette méthodologie a été mise en place sur notre corpus de parole et sera détaillée dans le chapitre 4.

2.2.3 L'enveloppe du signal pour la mesure du tempo

L'un des écueils majeur des méthodologies précédentes provient du fait qu'elles utilisent comme matériau de base à leurs analyses une segmentation (forward-backward ou détection d'onsets). Cela implique donc une représentation très simplifiée du signal dont la majorité des informations est perdue. Plusieurs auteurs se sont alors intéressés à l'utilisation d'autres méthodes pour fournir une représentation plus complète à la fonction de périodicité (auto-corrélation ou TFD). Scheirer (1998) a été le premier à remplacer la détection d'onsets par l'extraction de l'enveloppe d'amplitude du signal. L'enveloppe d'amplitude correspond à la courbe qui décrit les macro-variations d'amplitude d'un signal comme illustré sur la figure 2.7.

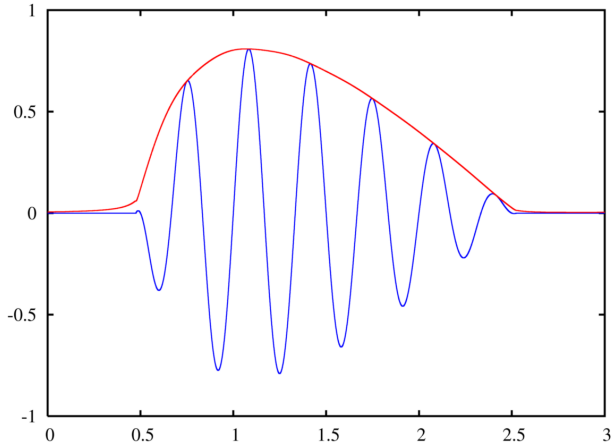

FIGURE 2.7 – Illustration d'un signal en bleu et de son enveloppe d'amplitude en rouge. Source : Omegatron CC BY-SA 3.0, `https://commons.wikimedia.org/w/index.php?curid=202296`

L'extraction de l'enveloppe peut se faire de nombreuses façons différentes. Scheirer (1998) utilise un ensemble de filtres passe-bandes afin d'obtenir plusieurs enveloppes avec des propriétés différentes où certaines sont davantage adaptées, par exemple, à des notes d'instruments. Tzanetakis et Cook (2002) quant à eux, utilisent un filtrage particulier en se basant sur des décompositions en ondelettes proposées par Daubechies (1988). Les enveloppes obtenues sont alors transformées en valeurs absolues, puis un nouveau filtrage est effectué, cette fois-ci afin de ne conserver que les variations lentes d'amplitude.

En sommant les enveloppes ainsi obtenues via les filtrages, il est alors possible d'appliquer la même méthodologie qu'illustré précédemment (figure 2.3) avec l'application de la fonction de périodicité sur la somme d'enveloppes directement sans avoir à passer par une segmentation. Cette méthodologie est parfois appelée *Beat histogram*. Une illustration de ce processus est donnée sur la figure 2.8.

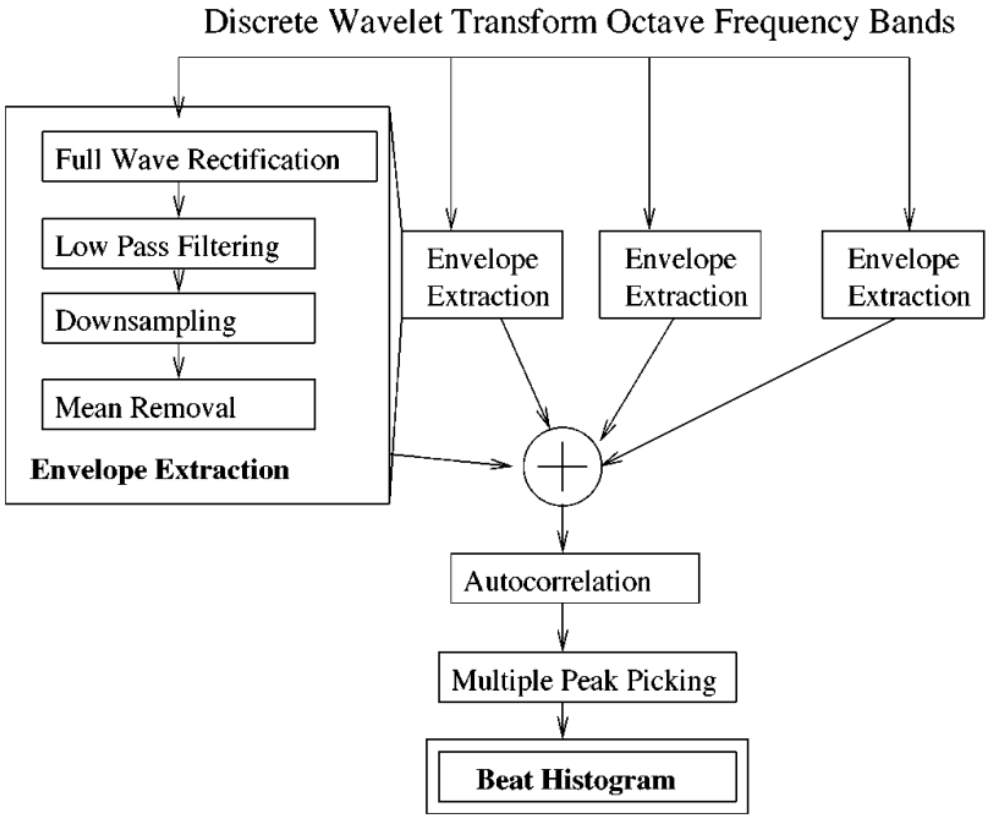

FIGURE 2.8 – Diagramme de calcul du *beat histogramme*. Figure tirée de Tzanetakis et Cook (2002, p.296)

2.3 Les spectres de modulations appliqués à la parole

Les méthodes décrites ci-dessus utilisées pour mesurer le tempo en musicologie ont des caractéristiques pertinentes pouvant être transposées à l'étude du rythme de la parole. C'est le cas notamment des modélisations basées sur l'extraction de l'enveloppe du signal en combinaison avec une analyse fréquentielle. En supposant que ce type de modélisations pouvait partiellement rendre compte du rythme de la parole, plusieurs auteurs ont adapté ces techniques pour les rendre pertinentes dans le traitement de la parole.

2.3.1 Spectre d'amplitude

Comme nous l'avons vu dans le premier chapitre, le rythme de la parole n'est pas un phénomène strictement périodique contrairement à la musique. Les régularités dans le flot de parole peuvent apparaître à plusieurs niveaux prosodiques plus ou moins larges. L'utilisation d'une méthode automatique prenant en compte les régularités du signal à différents niveaux (différentes fréquences) est donc une caractéristique très importante.

Tilsen et Johnson (2008) ont tout d'abord essayé de modéliser le rythme de la parole via une méthode proche de celle décrite dans la figure 2.8 en réalisant une extraction de l'enveloppe d'amplitude du signal initialement proposé par Cummins et Port (1998b). L'enveloppe est extraite en appliquant un filtre passe-bande entre 300 et 1300 Hertz qui selon Cummins et Port (1998b) est un filtrage permettant de supprimer l'information des fricatives (/f/, /s/ et /ʃ/) tout en conservant l'information des premiers formants des voyelles (nous reviendrons sur cet aspect que nous remettons en cause dans la section 4.2.1). Ce filtrage est alors suivi d'un passage en valeurs absolue et d'une soustraction de sa moyenne avant d'effectuer un filtrage passe bas (filtre de Butterworth (Butterworth et collab., 1930)) à 10 Hertz. Un exemple d'extraction d'enveloppe est donné dans la figure 2.9.

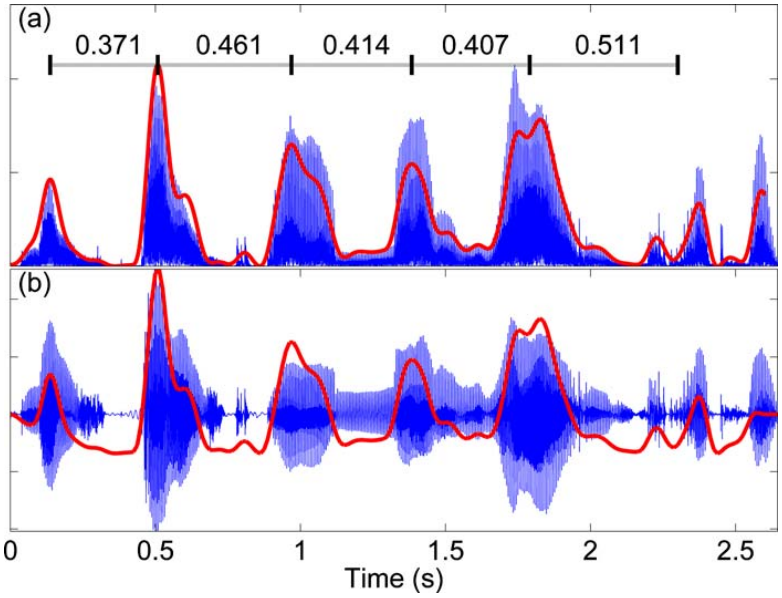

FIGURE 2.9 – Exemple de calcul de l'enveloppe (en rouge) d'un signal de parole. La phrase prononcée est "...category of Forrest Gump because Forrest Gump was great guy". La partie haute (a) montre en bleu la signal en valeurs absolues à partir duquel est calculée l'enveloppe (en rouge). En bas (b), l'enveloppe (à laquelle on a soustrait sa moyenne) est superposée au signal initial. Figure extraite de Tilsen et Johnson (2008, p.35)

Ce filtrage permettrait alors d'obtenir une enveloppe où les pics coïncideraient approximativement avec les onsets perçus des voyelles. À partir de cette enveloppe, il est alors possible d'étudier les différentes fréquences qui la composent en utilisant une fonction de périodicité comme la transformée de Fourier comme sur la figure 2.10 où l'on peut ainsi apercevoir un pic d'amplitude aux alentours de 2,2 Hz. L'apparition de ce pic signifie qu'une forme dans l'enveloppe du signal se répète de façon assez régulière à une fréquence de 2,2 apparitions par secondes.

Autrement dit, cette forme apparaît en moyenne toutes les 450 millisecondes environ dans l'enveloppe. En observant l'enveloppe correspondant en figure 2.9, il est alors

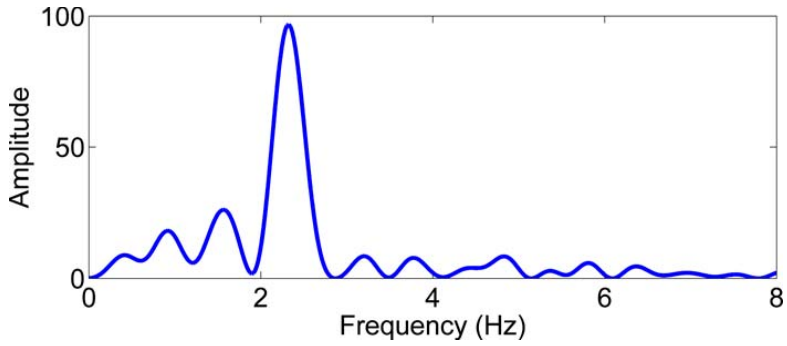

FIGURE 2.10 – Exemple de calcul de Transformée de Fourier de l'enveloppe correspondante au signal de parole de la Figure 2.9. Figure extraite de Tilsen et Johnson (2008, p.36)

possible de constater que la durée entre chaque pic est en moyenne de 435 ms ce qui correspond approximativement au pic du spectre d'enveloppe.

Cette méthodologie a par la suite été réutilisée et adaptée par plusieurs auteurs. Gibbon (2021) par exemple se base sur une méthode quasi-similaire, avec cependant une extraction de l'enveloppe d'amplitude différente basée sur le calcul maximum du signal sur une fenêtre glissante. D. Gibbon utilise alors le spectre de modulation d'amplitude dans le but de localiser des "formants" du rythme. Ces "formants" sont alors simplement les zones fréquentielles dans lesquelles se trouvent un pic d'amplitude du spectre. Afin d'obtenir une représentation des relations hiérarchiques entre les différentes zones de rythmes, Gibbon (2021) réalise également une classification hiérarchique ascendante basée sur les fréquences des pics afin de regrouper les zones de fréquences les plus proches. Cette méthodologie présente cependant quelques limites. Tout d'abord, la méthode d'extraction d'enveloppe n'est pas la plus pertinente et représente relativement mal les onsets des syllabes (MacIntyre et collab., 2022), de plus il étudie également un nombre fixe de pics de fréquences ce qui peut impliquer de considérer des pics de très faible amplitude comme pertinents.

L'utilisation de la Transformée de Fourier pour extraire les régularité de l'enveloppe est la méthode la plus couramment utilisée, mais il existe d'autres méthodes comme l'auto-corrélation comme nous l'avons décrit dans la section 2.2.1 ou encore la décomposition empirique du mode de l'enveloppe (Huang et collab., 1998; Tilsen et Arvaniti, 2013) qui se base sur une transformée de Hilbert afin d'extraire séparément les variations d'amplitudes liées aux syllabes et celles liées aux unités prosodiques supérieures.

2.3.2 Spectre de modulations de fréquences

Le spectre d'amplitude est donc un outil intéressant, mais il présente également des limites. Notamment, comme nous l'avons vu dans le chapitre précédent (section 1.1.6), l'amplitude (l'intensité) du signal n'est pas un paramètre très utilisé dans le cadre de l'accentuation des syllabes. Ce serait davantage la fréquence fondamentale qui jouerait

un rôle majeur dans la description de l'accentuation. Cela a donc mené à étudier un autre type de modulations : les modulations de fréquences. En effet, le signal peut varier à la fois au niveau de son amplitude, mais également au niveau des fréquences misent en jeu. L'exemple le plus simple correspondant aux modulations de fréquences est la courbe de fréquence fondamentale (f_0). Cette courbe montre l'évolution dans le temps de la f_0, mais les autres fréquences issues des vibrations des cordes vocales (harmoniques) et de la résonance du conduit vocal (les formants) sont complètement masquées.

Dans leur étude, Varnet et collab. (2017) font la distinction entre deux types de modulations de fréquences : le F_0M *spectrum* et le FM *spectrum*. Le spectre de f_0 consiste simplement à extraire la fréquence fondamentale du signal et d'y appliquer une fonction de périodicité. Cependant, le signal de parole n'étant pas constamment voisé, la f_0 extraite n'est pas continue. Ils utilisent alors une fonction de périodicité applicable à des signaux discontinus (ce qui n'est pas le cas de la TFD) appelée le *Root Lomb Periodogram*. Gibbon (2021) dans son modèle du rythme utilise également le spectre de f_0 en extrayant une courbe de f_0 à l'aide de l'algorithme AMDF (*Average Magnitude Difference Function*) proposé par Ross et collab. (1974). Il utilise alors une transformée de Fourier directement sur cette courbe en mettant les valeurs de f_0 dans les zones non-voisées à la valeur médiane de la f_0 globale.

Le calcul du FM *spectrum* quant à lui ressemble en partie à l'extraction du spectre d'amplitude. Le signal est également filtré mais plusieurs fois avec des filtres passe-bandes différents afin d'obtenir des informations sur toutes les bandes fréquentielles pertinentes. Par la suite, l'extraction de l'enveloppe de fréquence est obtenue en calculant la dérivée de la phase instantanée (*unwrapped instantaneous phase*) (Sheft et collab., 2008, 2012; Varnet et collab., 2017). Cette méthodologie génère cependant également des aberrations dans les fréquences instantanées au niveau des zones de silences. Un filtrage de la courbe est donc nécessaire en utilisant un filtre passe-bas par exemple ou encore en supprimant les informations dans les zones de silences et en utilisant le *Root Lomb Periodogram*.

2.3.3 Modulations spectro-temporelles

Maintenant que nous avons vu les deux types principaux de modulations (modulations de fréquences et d'amplitudes), il est important de mentionner une modélisation qui intègre ces deux informations, les modulations spectro-temporelles (MPS) (Chi et collab., 1999; Singh et Theunissen, 2003). L'extraction des modulations spectro-temporelles se déroule comme suit : 1) Extraction d'une représentation temps-fréquence du signal (spectrogramme) via une TFD sur fenêtre glissante et 2) application d'une Transformée de Fourier en deux dimensions sur ce spectrogramme. Ainsi, cette représentation permet de trouver des formes périodiques dans le spectrogramme de parole. Un exemple de MPS est donné sur la figure 2.11.

Dans cet exemple, il est possible d'observer que le spectrogramme de base est composé d'éléments simples (rayures rouges et bleues sur la figure). Chacune de ces formes

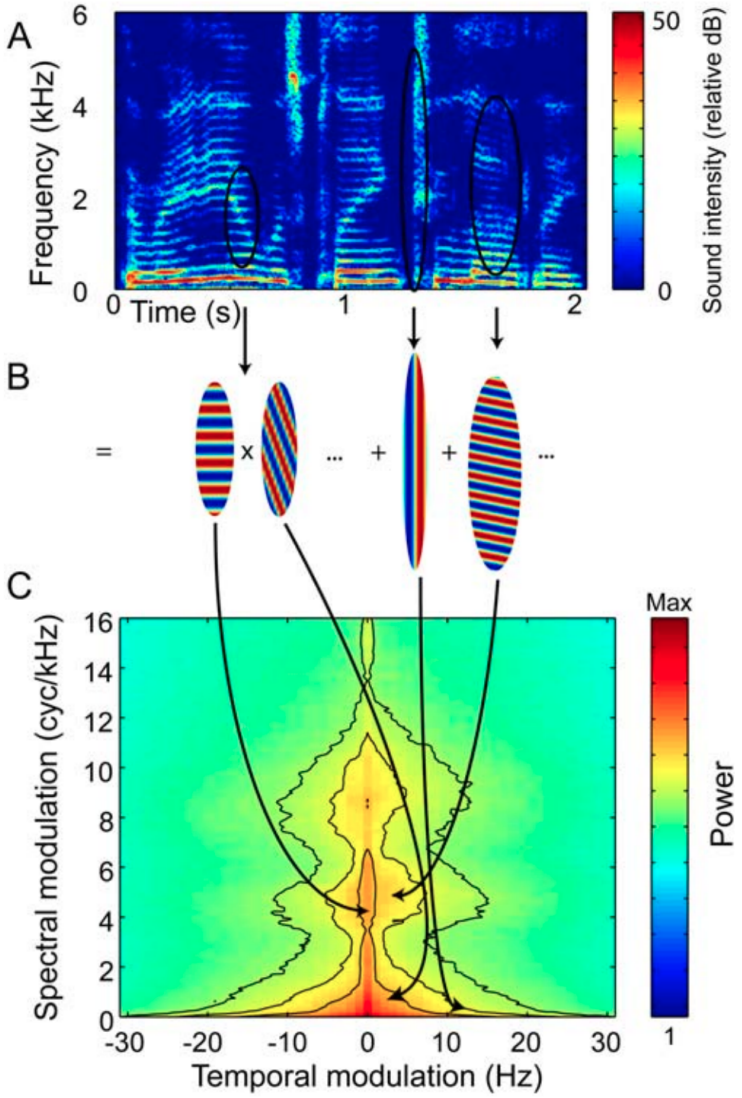

FIGURE 2.11 – Illustration de modulations spectro-temporelles sur un signal de parole lue ("*The radio was playing too loudly*"). Le spectrogramme du signal est donné en haut (A), au milieu (B), les formes caractéristiques qui forment le spectrogramme sont indiquées et en bas, le spectre des modulations spectro-temporelles est affiché. Figure extraite de Elliott et Theunissen (2009, p.2)

correspond à un endroit précis de la figure des modulations spectro-temporelles. Ainsi, la réalisation d'une consonne occlusive non-voisée (/p/, /t/, /k/) forme sur le spectrogramme une "rupture" visible où l'occlusion qui ne génère aucune énergie est suivie d'une explosion qui engendre une hausse d'énergie du spectre. Ce genre de réalisation peut donc être en grande partie modélisé par un motif formé de deux bandes verticales, respectivement de faible et haute énergie (troisième motif dans la partie B de la figure 2.11). La présence de ce genre de formes dans le spectrogramme implique alors un renforcement de l'énergie dans la représentation MPS au niveau de l'axe des abscisses (ici à environ 10 Hz) qui correspond aux modulations temporelles semblables en partie

aux modulations d'amplitudes. Une forte énergie autour de 10 Hz indique donc une forte présence d'occlusives d'environ 100 ms (correspondant à 10 Hz). D'autre part, la réalisation d'une voyelle stable en fréquence et en amplitude génère dans le spectrogramme des bandes d'énergies espacées régulièrement (harmoniques et formants) par rapport à l'axe des ordonnées. Cela se traduit par la première forme de la partie B de la figure composée de bandes horizontales plus ou moins espacées. Cette production ne génère donc pas de variation au niveau temporel étant donné que la voyelle est tenue à amplitude constante. En revanche, elle génère un ensemble de résonances à de nombreuses fréquences. Cela se traduit sur la figure MPS par une forte énergie au niveau de l'axe des ordonnées ici aux alentours de 5 cycles par kHz. La notation cyc/kHz indique alors le nombre de bandes d'énergies (formants / harmoniques) moyen dans un intervalle de 1000 Hz.

Les MPS ont souvent été utilisées afin d'évaluer quels sont les phénomènes dans la parole responsables de la compréhensibilité et l'intelligibilité (nous discuterons dans le chapitre 3 de la différence entre ces termes). Elliott et Theunissen (2009) par exemple ont modifié le signal de parole en supprimant des zones d'informations dans la représentation en MPS et ont montré que les modulations lentes de fréquences et amplitudes sont les zones les plus importantes. Les modulations d'amplitude entre 1 et 7 Hz étaient les zones les plus critiques pour assurer une bonne compréhensibilité, ainsi que les modulations de fréquence inférieures à 1 cyc/kHz.

Bien qu'intéressante, cette représentation est très peu intuitive. Il est très difficile d'extraire des informations visuellement de ces représentations et la majorité des études utilisant les MPS isolent les information des modulations spectrales ou temporelles avant de les utiliser. Il semble donc être trop tôt pour pouvoir extraire des informations rythmiques précises à partir de cette représentation. Il est nécessaire de continuer à explorer cette méthode afin de comprendre plus en détail la contribution conjointe des modulations de fréquence et d'amplitude sur des corpus de parole.

2.4 L'étude du rythme de la parole pathologique

Plusieurs modélisations automatiques du rythme se dégagent donc au sein de la communauté scientifique. Le rythme a été utilisé pour l'identification des langues ou l'analyse de la musique, mais il peut également être étudié dans le cadre de la parole pathologique afin d'aider à la détection de maladies en utilisant des méthodes non invasives, ou encore afin de caractériser les particularités prosodiques de pathologies.

Cependant, la plupart des études se concentrent sur le débit de parole. De plus, une grande partie des recherches sur le rythme de la parole pathologique se portent sur l'étude de la dysarthrie. La dysarthrie est un trouble de la parole qui se manifeste par des troubles articulatoires suite à un dysfonctionnement du système nerveux. Les maladies responsables de ces troubles peuvent être multiples, mais les principales sont la Maladie de Parkinson (MDP), les tumeurs au cerveau ou encore la sclérose en plaques. Tout d'abord, l'un des paramètres du rythme les plus simples à analyser est

le débit de parole. La comparaison du débit de parole entre une personne atteinte de la MDP et une personne saine, a été réalisée par quelques auteurs. Par exemple, Novotný et collab. (2014) ont mesuré une baisse du débit de parole chez les patients MDP diagnostiqués récemment dans le cadre d'une tâche de répétition rapide de syllabes (diadococinésie). En revanche, dans le cadre de parole lue, cette baisse de débit n'est pas toujours observée et il est même possible d'avoir une augmentation du débit de parole suite au phénomène de *local rush* qui fait référence à une accélération brutale du débit sur de courts segments de parole (Bayestehtashk et collab., 2015). Ces irrégularités se retrouvent dans l'étude de la diadochocinésie où les sujets MDP ont tendance à produire un rythme plus irrégulier de syllabes que les sujets sains (Novotný et collab., 2014; Rusz et collab., 2016). Sur cette même tâche, Rusz et collab. (2011) ont observé des variations anormales de l'intensité du signal entre deux répétitions de syllabes successives. Skodda (2015) supposent alors que le débit de parole et la capacité à produire un débit régulier correspondent à deux domaines différents des capacités motrices. Pour davantage de détails quant aux paramètres autres que le rythme les plus utilisés afin de distinguer des personnes saines de personnes MDP, nous vous renvoyons vers Jeancolas et collab. (2016) qui a réalisé un état de l'art des indices vocaux pour le diagnostic précoce de la MDP.

Les modélisations rythmiques ont également été utilisées pour aider à la classification de maladies. Liss et collab. (2009) ont ainsi été capable de discriminer 4 types de dysarthries en utilisant uniquement des informations sur le débit de parole et des caractéristiques comme ΔV, $\%V$ ou encore $VarcoV$ (caractéristiques décrites en section 2.1). À partir de ces mesures, ils ont pu obtenir un score de classification d'environ 80% ce qui montre l'intérêt de ces mesures dans l'étude de pathologies. Maffia et collab. (2021) ont également pu montrer que le $\%V$ chez des personnes atteintes de la MDP à un stade précoce était significativement plus élevé que chez les personnes saines. Ces études se sont donc concentrées sur l'étude du rythme au travers des mesures de durées inter vocaliques. Selon nous, il serait davantage intéressant de s'intéresser aux autres aspects du rythmes notamment à des niveaux supérieurs à la syllabe. Des études de la sorte ont été réalisées par exemple via l'analyse des spectres d'enveloppe. Liss et collab. (2010) ont extrait des spectres d'enveloppes sur un corpus de lecture de texte de 43 locuteurs présentant 4 types de dysarthries. Les paramètres extraits étaient l'amplitude et la fréquence du pic maximal de l'EMS, l'énergie dans les bandes de fréquences entre [3-6] Hz, [0-4] Hz et [4-10] Hz ainsi que le ratio entre l´énergie [0-4] et [4-10] Hz. Ils ont alors obtenu des scores de 84% de classification en utilisant ces paramètres. Les paramètres les plus discriminants entre les pathologies sont l'énergie dans la bande de fréquence [4-10] Hz qui est plus élevée chez les sujets contrôles. Également, le ratio d'énergie entre [0-4] et [4-10] Hz est plus faible élevé chez les patients. Leong et Goswami (2014a) ont aussi analysé l'EMS dans le cadre de la dyslexie développementale en utilisant une méthode analogue à celle de Liss et collab. (2010) en extrayant l'énergie dans des zones de fréquences particulières de l'EMS. Bien que ces paramètres soient intéressants à extraire, nous pensons que l'extraction d'énergie dans des bandes de fréquences fixes n'est pas toujours pertinente. En effet, dans le cas

où le débit de parole est variable entre les individus, il est possible que l'information située dans une bande comme [0-4] Hz ne contienne pas la même information d'une personne à une autre. Ainsi, une personne fortement impactée par sa maladie peut avoir un débit articulatoire très faible aux alentours de 2 Hz tandis qu'une personne saine pourra avoir un débit aux alentours de 5 Hz. Ainsi, l'énergie correspondant aux régularités syllabiques du patient se trouveront dans la bande [0-4] Hz alors que chez les sujet témoins, elle se trouvera dans la bande [4-10] Hz. Nous discuterons plus en détail de cette problématique et comment la résoudre dans la section 5.2.2.

2.5 Conclusion du chapitre

Dans ce chapitre, nous avons effectué un état de l'art des méthodes utilisées dans la littérature afin d'étudier le rythme de manière plus ou moins automatique. Nous avons ainsi vu dans un premier temps les différentes métriques du rythmes basées sur l'étude des régularités des segments vocaliques et consonantiques (Ramus et collab., 1999; Ling et collab., 2000; Dellwo, 2006) normalisés ou non en fonction du débit de parole. Bien que ces métriques permettent de modéliser des langues en se basant uniquement sur des régularités d'alternances consones/voyelles, il nous semble difficile de considérer qu'elles proposent une modélisation exhaustive du rythme. En effet, comme nous l'avons vu dans le chapitre 1, le rythme est un processus hiérarchique complexe qui ne peut pas être modélisé simplement au niveau syllabique. C'est l'une des raisons pour laquelle d'autres modélisations rythmiques sont apparues avec par exemple le tempogramme (Grosche et collab., 2010; Le Coz, 2014) qui est fortement inspiré des études faites sur l'estimation et le suivi du tempo en musique. Ce genre de modélisation permet d'intégrer la notion de hiérarchie en analysant le rythme à différentes fréquences simultanément. Ce principe très prometteur fait également face a des limites principalement au niveau du matériau de base utilisé pour calculer le tempogramme qui se base sur une segmentation du signal en unités très courtes qui perd une grande partie de l'information sur les modulations du signal de parole. Des modèles basés sur l'étude des régularités de modulations d'amplitude (Tilsen et Arvaniti, 2013; Varnet et collab., 2017) et de fréquence (Sheft et collab., 2008, 2012) de la parole ont alors émergé et permettent selon nous d'obtenir une modélisation satisfaisante du rythme de la parole. En effet, ils permettent de modéliser le côté hiérarchique de la parole, tout en conservant les modulations lentes du signal qui peuvent nous donner des indices sur l'accentuation des syllabes. Pour finir, ces modèles sont totalement automatiques car ils ne nécessitent pas de segmentation préalable du signal.

Du point de vue de l'étude du rythme dans la parole pathologique, peu d'auteurs se sont penchés sur le sujet, mais il semble tout de même que l'utilisation des modélisations cités précédemment puissent aider à la détection de certaines pathologies (Liss et collab., 2009, 2010; Leong et Goswami, 2014a; Maffia et collab., 2021). Cela nous conforte donc dans l'idée que le rythme peut être étudié dans certaines pathologies comme les dysarthries ou la MDP. A notre connaissance, aucune étude sur le rythme

n'a cependant concerné les cancers VADS. La seule étude en relation avec ce sujet est celle de Nocaudie et collab. (2018) dans laquelle les auteurs essayent de mettre en relation les fonctions prosodiques des patients avec leur compréhensibilité en réalisant une expérience de perception via des auditeurs naïfs. Cette étude semble suggérer que les patients auraient conservé voire amélioré leur prosodie afin de compenser leurs déficiences au niveau articulatoire. La modélisation automatique du rythme de la parole de patients atteints de cancers VADS pourrait donc potentiellement permettre de caractériser ces compensations. Cette caractérisation permettrait par exemple de vérifier si ces compensations prosodiques améliorent l'intelligibilité des patients.

3

Corpus et annotations

Sommaire

Dans les chapitres précédents, nous avons pu exposer notre définition du rythme d'un point de vue linguistique (chapitre 1) et présenter différentes modélisations automatiques du rythme (chapitre 2). Dans ce chapitre, nous allons présenter les deux corpus de parole sur lesquels nous avons pu effectuer différentes analyses qui seront exposées dans les chapitres suivants. Un petit corpus de slam est utilisé pour éprouver les analyses automatiques que nous conduisons dans les chapitres suivant sur nos deux corpus de parole pathologique.

Les deux corpus de parole pathologique sont issus du projet RUGBI [3] (*Looking for Relevant linguistic Units to improve the intelliGiBIlity measurement of speech production disorders*) financé par l'Agence Nationale de la Recherche (ANR). Il comprend des enregistrements audio accompagnés de méta-données sur les personnes enregistrées. Le corpus est composé de trois groupes :

— Des personnes atteintes de cancer de la cavité buccale et/ou du pharynx ayant

3. https://www.irit.fr/rugbi/

reçu une chimiothérapie et/ou une radiothérapie et/ou une chirurgie. Nous utiliserons simplement la dénomination "patients cancer" pour référer à ce groupe de personnes.

— Des patients atteints par la Maladie de Parkinson (MDP)

— Des sujets témoins sans troubles de la parole connus

Différentes tâches de production orale ont été enregistrées pour chaque locuteur afin de pouvoir au mieux évaluer les troubles dans différentes situations de communication. En plus des enregistrements audio, un ensemble de métadonnées a été récolté permettant d'avoir des informations sur l'âge, le type de tumeur (pour les patients cancers) ou encore l'âge de la maladie.

Ces enregistrements ont également été évalués par un jury de six experts qui ont jugé l'intelligibilité et la sévérité de chacun des locuteurs en produisant des scores basés sur plusieurs critères (Woisard et Lepage, 2010).

3.1 Description des corpus de parole

3.1.1 Corpus de slam

Tout d'abord, nous allons exposer les caractéristiques du corpus de parole slamée que nous avons eu à notre disposition. Ce corpus de slam nous a été gracieusement fourni par Anne-Catherine Simon de l'Université catholique de Louvain. C'est en réalité un extrait d'un corpus plus large. Ce sous-ensemble est constitué de quatre enregistrements audio de slam tiré de (Simon, 2020). Il contient des types de slam différents qui présentent des particularités rythmiques diverses. La Table 3.1 fournit une description des quatre enregistrements à notre disposition.

TABLE 3.1 – Description des fichiers du corpus de Slam. Tableau adapté de (Simon, 2020, p.146).

Titre	Auteur-interprète	Durée	Pauses	Débit	Syllabes accentuées
Slam	Petit Poussin	2 m 35 s	31%	4,0	19%
Dandy	L'Ami Terrien	1 m 47 s	7%	4,4	10%
Métastases	Skash	3 m 08 s	35%	4,4	22%
Une rencontre	NK	2 m 35 s	16%	4,1	31%

Parmi ces enregistrements, les quatre fichiers contiennent une majorité de passages dont le style de parole est dit *élaboré*. Cet adjectif appelé un phonostyle décrit la façon dont le slam est prononcé. Le style élaboré signifie que le texte est marqué par des rimes et que la production orale est scandée. La scansion consiste à prononcer le texte de façon à détacher les syllabes ou des groupes de mots. Ceci de sorte à avoir un rythme régulier qui "génère une pulsation rythmique entendue par l'auditoire" (Simon, 2020, p.143). Dans le cadre de cette thèse, nous nous sommes donc intéressés à ce style de parole car il produit les régularités rythmiques les plus fortes.

En plus des enregistrements, une annotation prosodique de ces fichiers a été réalisée par Anne-Catherine Simon. Cette annotation prosodique créée sur le logiciel Praat

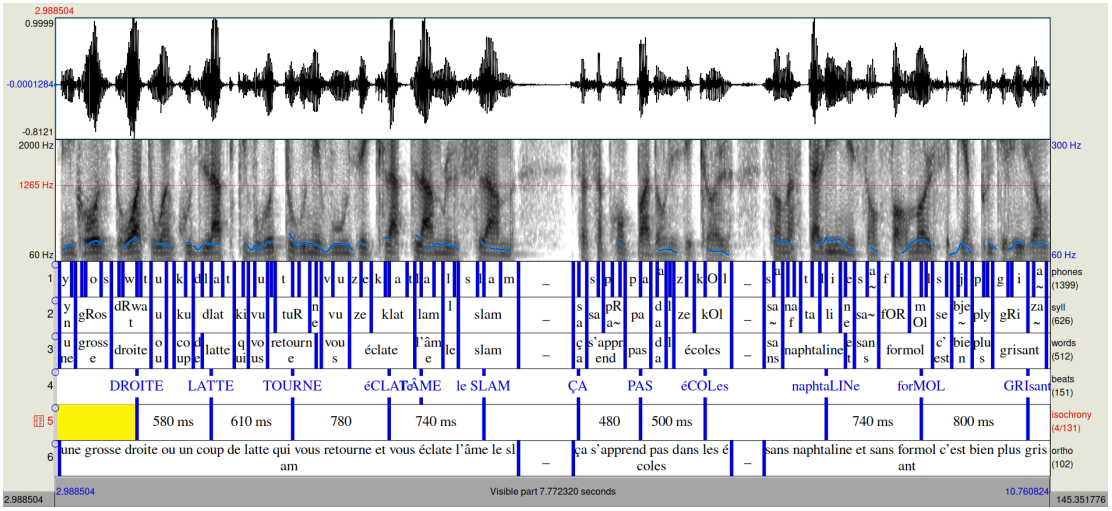

FIGURE 3.1 – Extrait du titre "Slam" avec une analyse des durées inter-accentuelles. Analyses prosodiques réalisées par Anne Catherine Simon (Simon, 2020).

(Boersma, 2001) nous fournit un ensemble d'informations visibles sur la figure 3.1. Parmi ces annotations, nous avons donc un alignement en phones, syllabes et en mots. À cela s'ajoute un repérage des beats (syllabes accentuées) et de leurs durées relatives. Cela nous permettra dans le chapitre suivant de vérifier que nos modèles parviennent bien à modéliser le rythme dominant d'un enregistrement audio. En effet, si les régularités rythmiques du slam ne sont pas détectées dans une modélisation automatique, il est alors très peu probable que des régularités dans de la parole lue classique ou pathologique soient correctement modélisées.

3.1.2 Corpus cancer VADS

Une fois nos modèles éprouvés sur les extraits de slam, nous pourrons donc les utiliser sur la parole pathologique. Le premier corpus de parole pathologique que nous allons détailler ici est le corpus de parole de personnes atteintes de cancer de la cavité buccale et/ou du pharynx. Ce corpus a été collecté dans le cadre du projet *Carcinologic Speech Severity Index*[4] (C2SI) financé par l'Institut National du Cancer (INCa). L'objectif de ce projet était de créer une base de données de voix françaises de personnes témoins et de patients atteints de cancers ORL et de valider des indices de mesure de sévérité et d'intelligibilité afin d'obtenir une évaluation objective de la qualité de vie des personnes atteintes de ce type de troubles de la parole.

Le corpus est composé d'un total de 113 locuteurs dont 87 patients atteints de cancers et 26 témoins. Plusieurs critères ont été vérifiés avant d'inclure les patients, à savoir que les patients devaient avoir reçu un traitement par chimiothérapie et/ou radiothérapie et/ou chirurgie depuis plus de 6 mois. Cette durée minimal permet alors d'être sûr que la dégradation au niveau de la parole est stable et que ces troubles

4. https://www.irit.fr/SAMOVA/site/projects/previous/c2si/

impactent la vie quotidienne des patients. Les patients présentant d'autres troubles connus de la parole indépendamment de leur pathologie n'ont pas été inclus. De même pour les personnes atteintes de troubles ne leur permettant pas de réaliser certaines tâches comme par exemple des troubles visuels qui ne permettraient pas de lire un texte convenablement.

L'âge moyen des locuteurs contrôles est de 55,7 ans (± 12,8) tandis que celui des patients est de 65,8 ans (± 9,5). L'âge entre ces deux groupes est significativement différents (p-valeur du test de Mann-Whitney inférieure à 0,02). Ceci est observable sur la figure 3.2 qui nous montre que le groupe contrôle est composé de 2 groupes d'âges différents, un premier autour de 40 ans et un second autour de 60 ans.

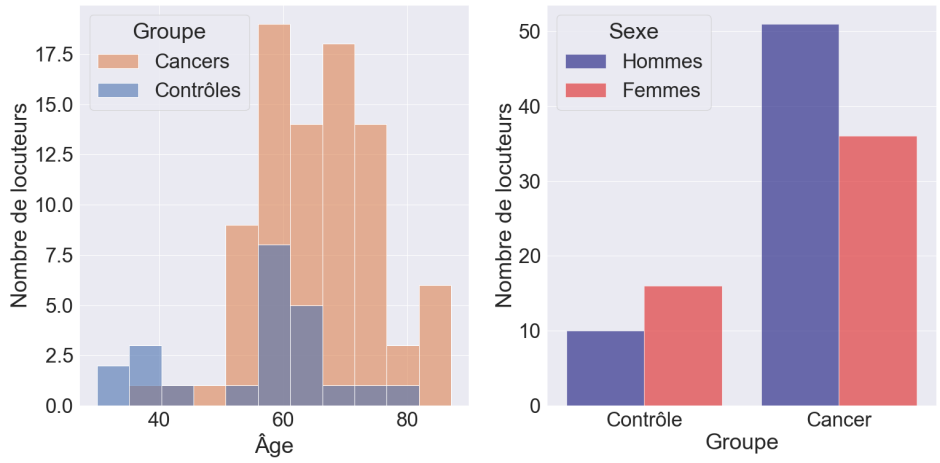

FIGURE 3.2 – Répartition de l'âge et du sexe des locuteurs dans le corpus en fonction du groupe (contrôle ou cancer). À gauche, la distribution de l'âge des locuteurs est affichée est fonction du groupe de patient. À droite, la répartition hommes/femmes est également affichée en fonction du groupe des locuteurs.

Au niveau du sexe des personnes enregistrées, le corpus est composé de 46% de femmes au total. Cependant, la proportion hommes/femmes dans le groupe cancer n'est pas équivalente à celle dans le groupe contrôle avec 41% de femmes dans le groupe cancer contre 62% dans le groupe contrôle. Cela est illustré dans la partie droite de de la figure 3.2.

Les enregistrements ont été effectués au Centre Hospitalier Universitaire de Toulouse au sein d'une chambre anéchoïque. Un microphone Neumann TLM 102 a été utilisé et les fichiers sont échantillonnés à 48 000 Hz. Une dizaine de tâches comme de la lecture de texte ou une description d'image ont été réalisées par les patients, mais tous n'ont cependant pas réalisé l'ensemble des tâches. Une description plus détaillée de ce corpus, avec la description de chaque tâche ainsi que des informations médicales sur les locuteurs est disponible dans l'étude Woisard et collab. (2021).

Les troubles de la paroles liés aux cancers oropharyngés et leurs traitements n'ont

pas été beaucoup étudié dans la littérature. En effet, les seules études à notre connaissance qui traitent de cette thématique sont celles liées au projet RUGBI auquel cette thèse est rattachée ainsi que le projet précédent C2SI. Ces recherches portent donc majoritairement sur la prédiction automatique de l'intelligibilité à partir de la parole.

Ma thèse s'inscrit dans un processus d'analyse de la prosodie et plus particulièrement du rythme de la parole dans le cadre de ces cancers, ce qui n'a jamais encore été réalisé. Les particularités prosodiques de ces patients sont donc encore inconnues bien que cette maladie induise principalement des troubles respiratoires et de la déglutition ayant potentiellement un impact sur le rythme, la fluence et donc l'intelligibilité de la parole des patients. L'étude du rythme de la parole semble donc être un élément pertinent pour connaître davantage ces troubles.

3.1.3 Corpus Parkinson

En plus du corpus cancer, un corpus de parole parkinsonnienne a été intégré au projet ANR RUGBI au cours de cette thèse. Ce corpus a été constitué dans le Service de Neurologie du Centre Hospitalier du Pays d'Aix à Aix-en-Provence et provient de la base de données AHN (Ghio et collab., 2012). Il contient au total 316 locuteurs 205 patients parkinsoniens et 111 sujets témoins. La répartition en âge est davantage équilibré entre les deux groupes avec une moyenne d'âge de 66,5 ans ($\pm$=9,6) pour les patients parkinsoniens et de 62,7 ans pour les témoins ($\pm$=11,0). La différence des moyennes est significative entre les deux groupes mais seulement 4 ans les différencient ce qui ne devrait pas avoir un fort impact en termes de vieillissement de la voix et notamment au niveau prosodique (Hixon et collab., 2008). La répartition de l'âge des patients est illustrée sur la partie gauche de la figure 3.3.

La répartition hommes/femmes est assez inégale entre les groupes avec une majorité d'hommes chez les personnes atteintes de la maladie de Parkinson (67% d'hommes). Ce déséquilibre peut être justifié par le fait que la maladie de Parkinson touche en moyenne 1,5 fois plus les hommes que les femmes d'après sur les données nationales de surveillance de la fréquence de la MDP entre 2010 et 2015 [5]). En revanche, chez les patients contrôles, les hommes ne représentent que 41% des personnes. Ce phénomène pourrait être expliqué par le choix des participants à l'étude car les conjoint(e)s des patients sont davantage susceptibles d'accepter de participer à ce genre d'études. La répartition du sexe des locuteurs est illustrée dans la partie droite de la figure 3.3.

L'impact de la MDP sur la prosodie des personnes a été étudié de façon sporadique avec une majeure partie des études s'intéressant à l'étude du débit de parole ou à la fréquence fondamentale. Encore aujourd'hui, les effets de cette maladie sur la f_0 ne sont pas parfaitement connus avec certaines études montrant que les patients présentent une f_0 inférieure à celle de sujets témoins (Meynadier et collab., 1999; Viallet et collab., 2000). En revanche, d'autres prédisent une hausse due à une rigidité du muscle tenseur des cordes vocales qui générerait une augmentation de la pression sous la glotte et donc une f_0 plus élevée (Goberman et Coelho, 2002). Cette hausse

5. https://solidarites-sante.gouv.fr/IMG/pdf/rapport-frequence-maladie-parkinson-france.pdf

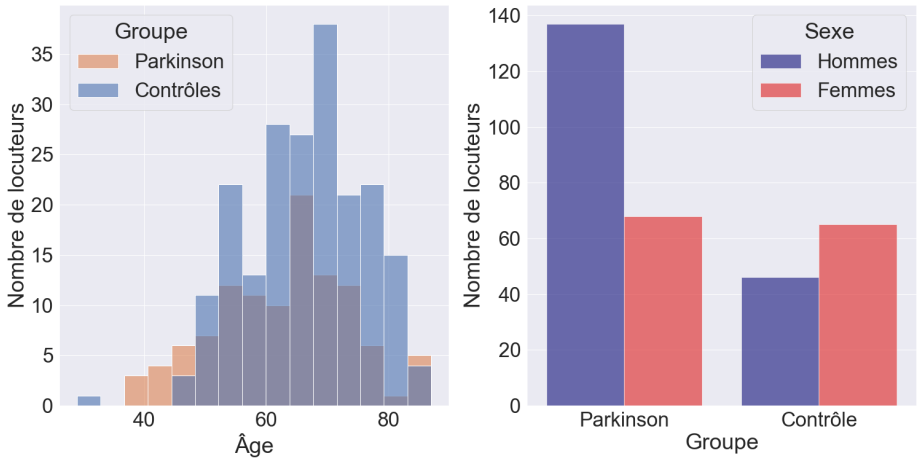

FIGURE 3.3 – Répartition de l'âge et du sexe des locuteurs dans le corpus en fonction du groupe (contrôle ou Parkinson). À gauche, la distribution de l'âge des locuteurs est affichée est fonction du groupe de patient. À droite, la répartition hommes/femmes est également affichée en fonction du groupe des locuteurs.

a pu été observée sur certains corpus (Metter et Hanson, 1986). La modification de hauteur de f_0 n'est donc pas un consensus global, en revanche la grande majorité des études sur le sujet ont pu constater une diminution des modulations de f_0 et donc une limitation des variations de leur voix.

Bien que la MDP ait des effets néfastes sur la parole, il existe des traitements médicamenteux à base de Levadopa (L-DOPA). Ce traitement permet une amélioration globale des troubles moteurs liés à la MDP. En revanche, l'effet du traitement sur la parole n'est pas encore très clair (Pinho et collab., 2018; Tykalova et collab., 2022). Une particularité intéressante du corpus parkinsonien est que les patients ont été enregistrés deux fois :

— Une première fois une heure après leur prise habituelle de L-DOPA (ON-DOPA).
— Une seconde fois sous sevrage médicamenteux de plus de 12h (OFF-DOPA).

Afin de clarifier l'effet qu'ont la MDP et son traitement sur la f_0, nous avons réalisé une comparaison des mesures de f_0 entre les différents groupes (Contrôles, ON-DOPA et OFF-DOPA). Cette étude a été publiée dans la conférence francophone des Journées d'Études de la Parole (JEP) 2022 (Vaysse et collab., 2022b). Nous avons ainsi observé une très légère hausse de f_0 chez les patients atteints de MDP (principalement chez les hommes), mais plus étonnamment une nouvelle hausse (très légère) chez les patients ON-DOPA par rapport aux patients OFF-DOPA. Également, nous avons pu observer une baisse significative des modulations de f_0 chez les patients MDP, mais n'avons pas trouvé d'effet significatif du traitement par L-DOPA en faveur d'un regain de modulations. Les résultats sur les modulations de f_0 sont visibles sur la figure 3.4.

L'effet du traitement par L-DOPA sur la f_0 ne semble pas très marqué sur notre

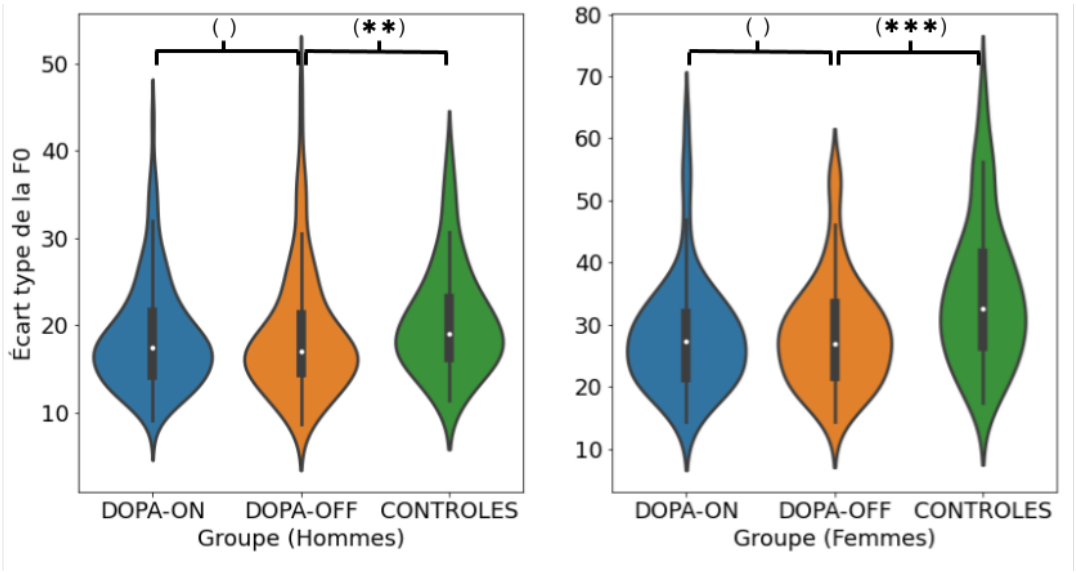

FIGURE 3.4 – Répartition de l'écart-type de la F0 en fonction du sexe et du groupe de locuteurs. Les (*) indiquent si une différence significative existent ou non entre les groupes : ( ) non significatif ; (**) p-valeur < 0,01 ; (***) p-valeur < 0,001. Figure extraite de Vaysse et collab. (2022b, p.312)

corpus tandis que la MDP en elle-même pourrait avoir une incidence négative sur la prosodie des patients.

3.1.4 La tâche de lecture de texte

Nos deux corpus (cancer et MDP) sont composés d'une dizaine de tâches de production orale différentes. Parmi elles par exemple, on retrouve une tâche où les locuteurs doivent lire à voix haute une liste de pseudo-mots (respectant les règles phonotactiques de la langue mais n'ayant pas de sens), une lecture de texte ou encore des tâches dites prosodiques. Les tâches prosodiques incluent notamment :

— Une tâche de modalités où les locuteurs lisent des phrases courtes isolées avec différentes intonations (question, affirmation, ordre), l'objectif étant de voir si les personnes parviennent à moduler leur fréquence fondamentale.

— Une tâche de focus dans laquelle ils devaient mettre en exergue un mot particulier dans une phrase via une réponse à une question (par exemple, question : *"Tu as vu un canard ou un cochon dans le jardin ?"* ; réponse : **"J'ai vu un CANARD dans le jardin"**)

— Une tâche de désambiguïsation syntaxique dans laquelle ils devaient prononcer une phrase contenant une ambiguïté comme par exemple : *"les baguettes et les croissants chauds"* dans laquelle la prosodie du locuteur devrait être différente si l'adjectif concerne les deux noms ou seulement le dernier.

Bien que ces trois tâches puissent être pertinentes pour évaluer les modulations de f_0 chez les locuteurs, il s'avère que la façon de lire des personnes joue un rôle trop

important pour pouvoir exploiter correctement ces exercices. En effet, de nombreux sujets témoins ont des difficultés pour réaliser ces tâches étant donné que les exemples sont très courts et peu naturels. L'exploitation de ces tâches pour l'analyse du rythme de la parole ne nous a donc pas semblé être pertinent.

Parmi les exercices restants, la plus propice selon nous pour évaluer la prosodie est la lecture de texte. En effet, nous pensons qu'il est nécessaire d'avoir un matériel suffisamment long afin de pouvoir étudier les variations sur plusieurs syntagmes. La description d'images, plus spontanée aurait également pu être un candidat, mais le fait de pouvoir comparer des signaux davantage similaires dans le cas de la lecture (même texte pour tous) nous semble être un avantage non-négligeable en faveur de la lecture de texte et un premier pas dans une étude exploratoire sur le rythme en parole pathologique. Le texte en question est un extrait du livre d'Alphonse Daudet, "la chèvre de Monsieur Seguin" (Daudet, 1870). Le passage en question est le suivant :

> Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon. Un beau matin, elles cassaient leur corde, s'en allaient dans la montagne, et là haut, le loup les mangeait. Ni les caresses de leur maitre, ni la peur du loup, rien ne les retenait. C'était parait-il des chèvres indépendantes voulant à tout prix le grand air et la liberté. (2)

Le texte est relativement long et composé de plusieurs phrases variées du point de vue de leur organisation. Bien que l'organisation prosodique d'une phrase dépende principalement du locuteur, une structuration prosodique potentielle des trois premières phrases est disponible en figure 3.5.

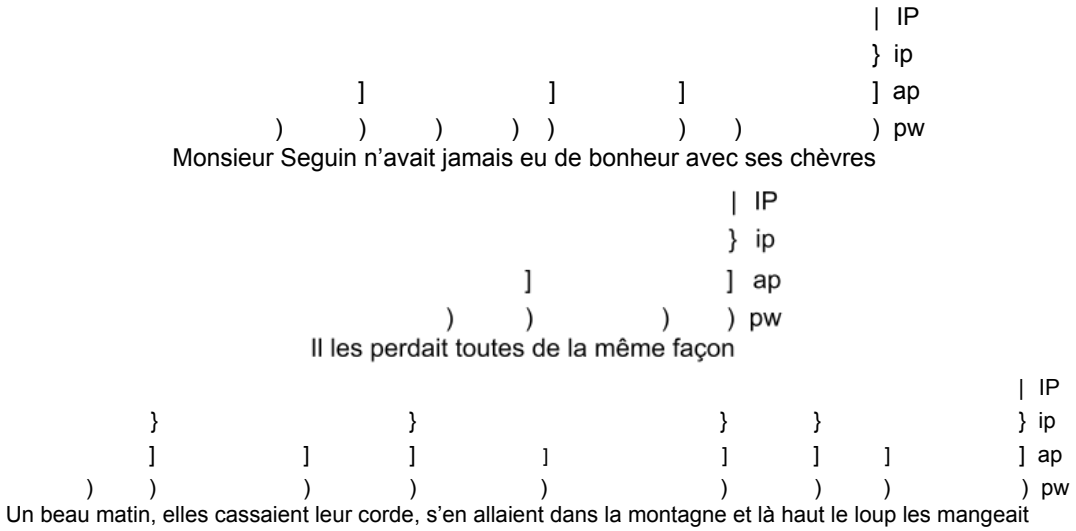

FIGURE 3.5 – Organisation prosodique potentielle des trois premières phrases de la lecture de texte. IP : Syntagme intonatif, ip : syntagme intermédiaire, ap : syntagme accentuel, pw : mot prosodique. Voir la section 1.1.3 pour une description détaillée des niveaux.

Les structures prosodiques potentielles de ces trois phrases sont relativement complémentaires, avec une première phrase composée d'une seule IP suffisamment longue, composée de plusieurs ip, tandis que la seconde est composée d'une seule IP qui est également une ip. La dernière phrase quant à elle est composée de quatre ip qui imposent un rythme avec une forme d'énumération qui devrait imposer une régularité à ce niveau. La quatrième phrase (qui n'est pas illustrée ici) ressemblent fortement au niveau prosodique à la troisième (Une énumération de trois ip).

Ces structures prosodiques ne sont que des propositions théoriques sujettes à variation en fonction du locuteur. Il est donc crucial d'avoir des informations sur la qualité de production des locuteurs au travers d'annotations expertes pour pouvoir évaluer efficacement la prosodie des patients.

3.1.5 Annotations cliniques perceptives

Afin d'obtenir un ensemble de mesures liées aux troubles de la parole des patients des deux corpus MDP et Cancer, un jury de six experts ORL a été mis en place afin de juger les productions orales des locuteurs. Parmi ces jugements, il a notamment été demandé aux juges d'évaluer la sévérité et l'intelligibilité des locuteurs. Ces notions d'intelligibilité et de sévérité ne sont pas toujours définies de la même façon chez tous les auteurs. Dans cette thèse, nous nous plaçons dans la logique de Pommée et collab. (2022) qui ont établi une définition de l'intelligibilité à partir d'une méthodologie Delphes, c'est à dire en consultant un grand nombre d'experts de sorte à produire une définition consensuelle. Dans cette étude, il considère l'intelligibilité comme suit :

Intelligibility refers to the reconstruction of an utterance at the acoustic–phonetic level, intelligibility-related information is thus carried by the acoustic signal (i.e., intelligibility focuses on signal-dependent information).This reconstruction is made possible both by the speaker's phonetic–acoustic production ability and by the listeners acoustic–phonetic decoding skills. (Pommée et collab., 2022, p.31)

"L'intelligibilité fait référence à la reconstruction d'un énoncé au niveau acoustico-phonétique, l'information liée à l'intelligibilité est donc portée par le signal acoustique (c.-à-d. l'intelligibilité se concentre sur l'information dépendante du signal). Cette reconstruction est rendue possible à la fois par les capacité de production accoustico-phonétique du locuteur et par les capacités de décodage acoustico-phonétique de l'auditeur".

La sévérité quant à elle se définit davantage comme une mesure de l'impact de la pathologie sur la production orale d'un locuteur. La sévérité est plus générale que l'intelligibilité et ne fait pas référence aux capacités de décodage acoustico-phonétique de l'auditeur (Woisard et Lepage, 2010).

En amont de l'évaluation de l'intelligibilité et de la sévérité, les six experts ont évalué le degré d'altération des quatre paramètres perceptifs :

— La qualité de la voix

— La résonance
— La prosodie
— La distorsion phonémique

Ces indices ont été évalués perceptivement sur une échelle discrète entre 0 (pas d'altération) et 3 (altération sévère). Aucune définitions de ces concepts n'ont été préalablement données aux évaluateurs. Le jugement de ces paramètres avait pour but d'aider le jury à évaluer la sévérité et l'intelligibilité des patients de façon plus objective. Les six juges experts ont par la suite produit une notation entre 0 et 10 pour ces deux scores (10 = aucun trouble, 0 = indice très dégradé) et enfin la moyenne des six scores obtenus a été calculée pour chaque locuteur. La répartition des scores d'intelligibilité et de sévérité sur le corpus de parole cancer est disponible dans la figure 3.6.

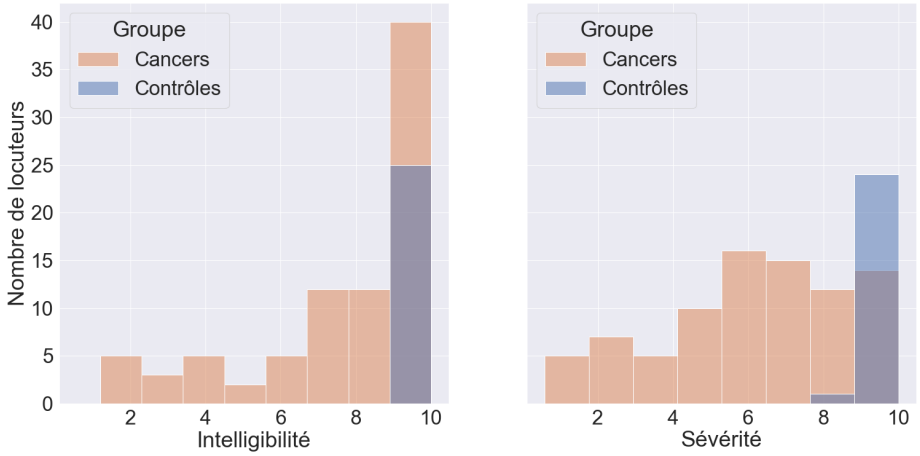

FIGURE 3.6 – Histogramme de l'intelligibilité (à gauche) et de la sévérité (à droite) pour les populations cancer (orange) et les sujets témoins (en bleu)

Sur cette figure, nous pouvons voir que les sujets témoins ont tous une intelligibilité supérieure à 9 et une sévérité supérieure à 8 (10 étant le maximum). Du côté des patients, un nombre important d'entre eux possèdent également un score d'intelligibilité élevé (52 locuteurs ont un score supérieur à 8). Du côté de la sévérité, les scores sont mieux répartis sur l'ensemble de la plage de scores possibles.

Concernant le corpus Parkinson, la répartition des scores d'intelligibilité et de sévérité en fonction du groupe de locuteur est donnée dans la figure 3.7.

Dans le cadre de ce corpus, les scores sont davantage élevés avec un score d'intelligibilité minimal de 5 et un score de sévérité minimum de 3,5. Il est intéressant de noter que la sévérité des locuteurs témoins est plus étalée que dans le cadre des témoins enregistrés dans le corpus cancer, cela peut s'expliquer par la moyenne d'âge plus élevée dans le corpus Parkinson. Au niveau d'un effet possible du traitement par L-DOPA sur ces scores, nous n'avons pas détecté de différences significatives pour ces scores entre les groupes ON-DOPA et OFF-DOPA en utilisant un test de Wilcoxon ($p_{intel} = 0,12$ et $p_{sev} = 0,16$). Cela pourrait donc montrer que dans notre corpus, l'effet

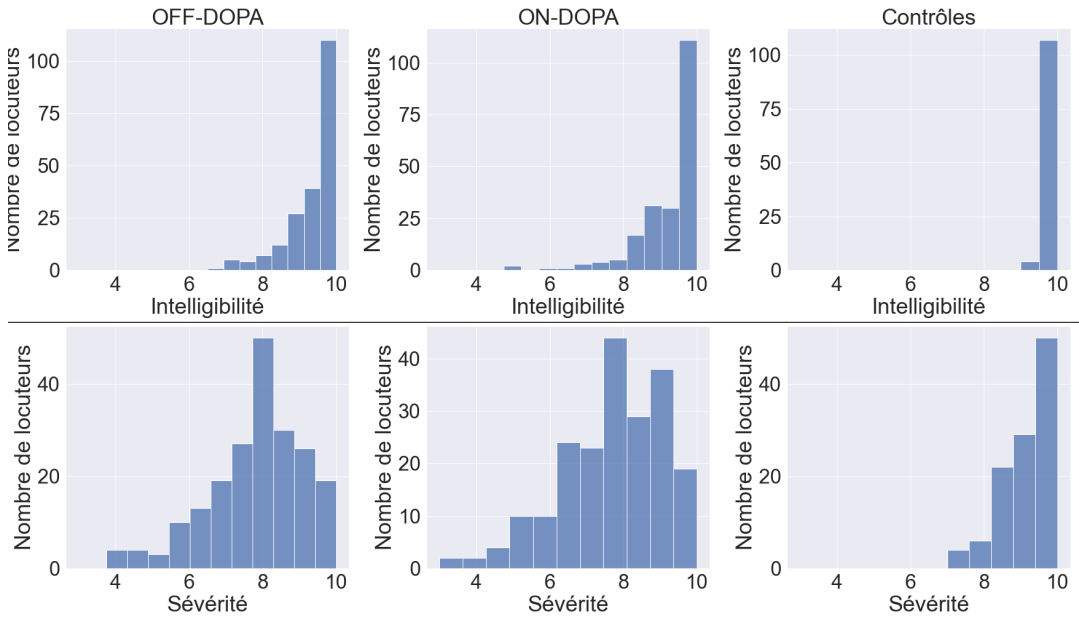

FIGURE 3.7 – Histogramme de l'intelligibilité (en haut) et de la sévérité (en bas) pour le corpus MDP dans le cas des différents groupes. À gauche, les patients sous sevrage médicamenteux, au milieu, les patients sous traitement et à droite les sujets contrôles

du traitement par L-DOPA ne montre pas d'amélioration significative des troubles de la parole des patients.

3.2 Annotations prosodiques

Les corpus cancer et Parkinson que nous avons à notre disposition incluent donc de nombreuses informations cliniques sur la qualité des productions orales des patients. En revanche, au niveau prosodique, seul le score subjectif de prosodie des experts est disponible. Cependant, sans définition claire de ce score de la part des cliniciens, il est difficile de prendre ce score comme référence afin de caractériser le rythme des personnes enregistrées. Un ensemble d'annotations plus ciblées ont donc été créées dans le but d'avoir une référence fiable quant aux caractéristiques prosodiques des locuteurs.

3.2.1 Annotation de la fréquence fondamentale

Dans un premier temps, nous avons vu dans les chapitres précédents que la f_0 joue un rôle important dans l'étude du rythme de la parole. Il existe de nombreux algorithmes permettant d'estimer automatiquement la fréquence fondamentale d'une personne à partir d'enregistrements audio. Cependant, ces algorithmes sont aujourd'hui très performants dans le cas de parole classique, mais n'ont que rarement été évalués sur de la parole pathologique (Parsa et Jamieson, 1999; Jang et collab., 2007). Afin de

savoir quel algorithme d'estimation de f_0 est le plus performant dans le cadre de nos corpus de parole pathologique, nous avons décidé de mesurer les performances d'une dizaine d'algorithmes sur nos données Vaysse et collab. (2022a). Une condition nécessaire pour pouvoir réaliser cette évaluation est d'avoir une f_0 de référence à laquelle comparer les estimations des algorithmes. Les fichiers n'ayant pas été enregistrés à l'aide d'un Electro-Glotto-Graph, il a été nécessaire d'annoter la f_0 manuellement.

Il aurait été trop coûteux en temps de réaliser cette annotation sur l'ensemble du corpus de parole pathologique. C'est pourquoi nous avons décidé de n'annoter qu'une partie du corpus. Nous avons donc sélectionné un ensemble de 24 locuteurs (8 témoins, 8 cancers, 8 MDP) afin d'annoter la f_0 de leur lecture tâche de lecture. Les locuteurs présentant une pathologie ont été sélectionnés en se basant sur l'indice de qualité de leur voix (section 3.1.5). Nous avons supposé que les personnes avec un indice de trouble vocalique élevé seraient plus susceptibles de présenter des troubles au niveau de leur f_0 qui pourraient induire en erreur les algorithmes de détection automatique.

Pour obtenir des valeurs de f_0 fiables, nous avons utilisé le logiciel Praat (Boersma, 2001) dans le but d'effectuer une première annotation automatique à l'aide d'un algorithme basé sur l'autocorrélation du signal. Nous avons par la suite corrigé cette annotation manuellement au niveau du signal en indiquant les limites de chaque période pour obtenir une fréquence fondamentale correspondant à la valeur réelle (voir figure 3.8). Une fois corrigée, la courbe de fréquence fondamentale a été extraite avec des valeurs toutes les 10 ms (voir figure 3.9).

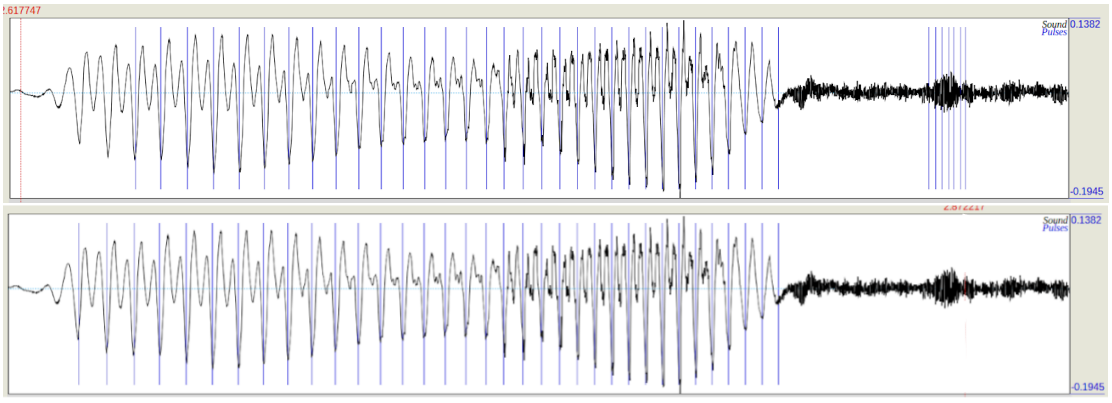

FIGURE 3.8 – Exemple de correction manuelle de la f_0 sur un enregistrement de personne atteinte de cancer. Le calcul automatique fourni par Praat est au dessus, notre correction manuelle au dessous.

La génération de ces valeurs corrigées de f_0 nous permet donc d'avoir une référence fiable dans le but de pouvoir évaluer les performances d'un ensemble d'algorithmes de détection automatique de f_0 afin de savoir quels sont les meilleurs algorithmes à utiliser dans le cadre de ces pathologies. Les résultats de cette étude seront décrits plus tard dans la section 4.4.1

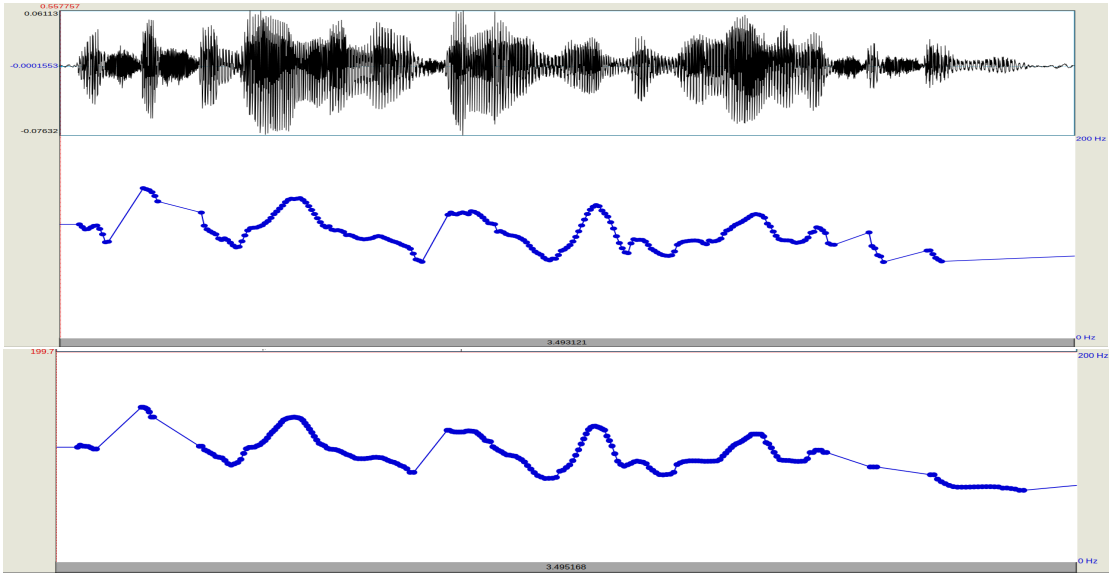

FIGURE 3.9 – Exemple de courbe de f_0 obtenue avant et après correction de l'algorithme de Praat sur la phrase « Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres» d'un sujet témoin.

3.2.2 Annotation de la structure prosodique

L'un des objectifs principaux de cette thèse est de caractériser le rythme de la parole pathologique. Pour cela, il est important de mettre en relation les modélisations automatiques présentées dans le chapitre 2 avec des annotations de la structuration prosodique pour savoir si ces modèles peuvent rendre compte efficacement des stratégies prosodiques des locuteurs. Pour cela, nous aurions pu utiliser le score perceptif d'évaluation de la prosodie présenté dans la section 3.1.5. Cependant, aucune définition de la prosodie n'a été proposée aux évaluateurs avant leur jugement. Il est donc difficile de savoir quels aspects prosodiques (f_0, rythme, fluence...) ont été pris en compte dans leurs jugements. De plus, l'échelle de notation discrète de 0 à 3 ne permet pas selon nous de rendre compte précisément du degré des troubles rythmiques. C'est la raison pour laquelle nous avons décidé d'annoter en plus les niveaux prosodiques dans des productions orales de lecture de texte. Les niveaux annotés sont la syllabe (syll), le mot prosodique (pw), le syntagme accentuel (ap), le syntagme intermédiaire (ip) et le syntagme intonatif (IP). La description de ces différents niveaux a été faite dans la section 1.1.3. Ces annotations supplémentaires nous permettront d'avoir une évaluation plus précise des troubles présents au niveau du rythme de la parole.

Nous avons voulu nous concentrer uniquement sur le corpus de parole cancer VADS qui selon nous présente des particularités prosodiques variées de par leurs scores de sévérité plus dispersés (Figure 3.6) par rapport aux patients atteints de la MDP (Figure 3.7). Nous avons donc sélectionné 10 patients atteints de cancer VADS et 10 locuteurs sains. Les 10 patients ont été choisis arbitrairement de sorte à avoir des personnes dont la prosodie nous semble caractéristique de ce que l'on peut retrouver

dans ce corpus. Ainsi, nous avons des personnes ne semblant pas avoir d'atteinte prosodique, ainsi que des locuteurs avec des troubles respiratoires et/ou articulatoires divers impliquant des stratégies de groupements prosodiques variées.

Nous avons personnellement annoté les niveaux de la syllabe et du mot en utilisant le module EasyAlign (Goldman, 2011) qui permet de réaliser un alignement forcé d'un texte sur un enregistrement audio dans le logiciel Praat. Nous avons par la suite corrigé manuellement cet alignement. Nous avons également généré automatiquement le niveau du mot prosodique à partir du niveau du mot en fusionnant les mots lexicaux et leur(s) clitique(s). L'annotation prosodique des autres niveaux a en revanche été réalisée par Corine Astésano[6]. Un exemple d'annotation prosodique d'une personne saine est présentée dans la figure 3.10.

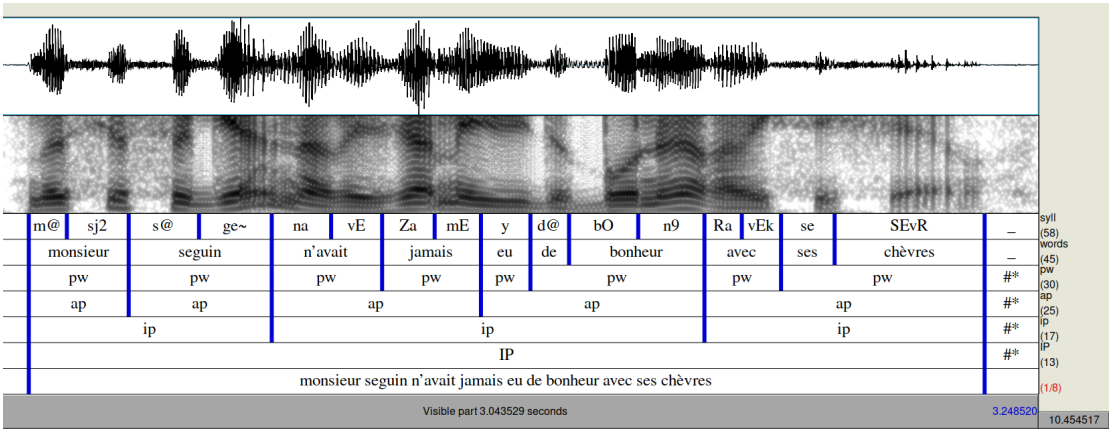

FIGURE 3.10 – Exemple d'annotation prosodique sur le logiciel Praat pour un locuteur témoin sur la première phrase de la lecture de texte. On retrouve la forme d'onde du signal en haut, son spectrogramme en dessous et enfin les différentes tires concernent les annotations aux niveaux syllabique (syll), pw, ap, ip et IP

Dans cet exemple, nous retrouvons approximativement la structure que nous avions proposé dans la figure 3.5 étant donné que la personne est un sujet témoin ne présentant pas de troubles particulier de la parole. En revanche, la production d'une personne atteinte d'un trouble sévère de la parole ne correspond plus du tout à notre structure comme illustré dans la figure 3.11 avec une personne dont le score perceptif de sévérité est de 1,8.

Cet ensemble d'annotations nous permettra alors de pouvoir vérifier si les modélisations automatiques sont capables ou non de rendre compte de la régularité prosodique à différents niveaux et de voir si les patients atteints de cancer présentent des stratégies de programmation de la parole différentes des sujets témoins.

3.2.3 Catégorisation libre de la prosodie des locuteurs

Grâce à la segmentation en niveaux prosodiques des locuteurs décrits dans la partie précédente, nous pourrons donc avoir des données quantitatives pour évaluer nos

6. https://lnpl.univ-tlse2.fr/accueil/membres/corine-astesano-1

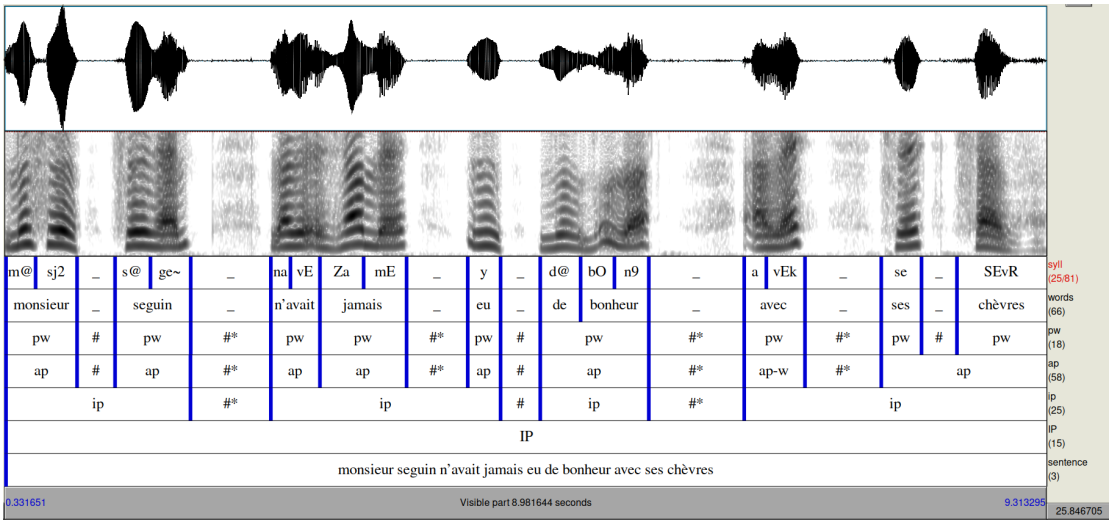

FIGURE 3.11 – Exemple d'annotation prosodique sur le logiciel Praat pour un locuteur atteint de cancer sur la première phrase de la lecture de texte. On retrouve la forme d'onde du signal en haut, son spectrogramme en dessous et enfin les différentes tires concernent les annotations aux niveaux syllabique (syll), pw, ap, ip et IP. Les pauses sont indiquées par les symboles # et les pauses respiratoires par #*

modélisations automatiques. En revanche, comme nous l'avons vu dans la section 1.2.3, la réalité du rythme ne se retrouve pas toujours uniquement dans les régularités acoustiques. La perception des auditeurs joue en effet un rôle majeur dans l'évaluation de la prosodie. C'est la raison pour laquelle nous avons voulu ajouter ce type d'information sur notre sélection de 20 locuteurs. Pour cela, nous avons choisi d'effectuer une catégorisation libre du rythme de ces locuteurs sur leur lecture de texte. Tout comme les annotations des niveaux prosodiques, cette catégorisation a été réalisée par Corine Astésano. Le but de cette tâche est d'écouter un par un chacun des 20 enregistrements de lecture et de les placer dans un plan en deux dimensions de sorte à placer côte à côte deux personnes dont le rythme perçu est similaire.

Cette catégorisation a été réalisée en écoutant un par un chaque locuteur et en le plaçant sur le plan. Pour chaque nouveau locuteur, l'enregistrement pouvait être écouté plusieurs fois et a été placé plus ou moins proche des personnes déjà placées. Les critères de placement étaient principalement focalisés sur le rythme de la parole en essayant de s'abstraire de la sévérité de la pathologie des patients. Une fois tous les locuteurs placés, une seconde écoute de chaque enregistrement a été effectuée de sorte à corriger leur placement. Le résultat de cette catégorisation est visible sur la figure 3.12. À la suite de cette catégorisation, nous avons exploré ce corpus de sorte à essayer de comprendre à posteriori à quoi les différents axes du plan correspondent. Nous nous sommes rendus compte que l'axe des ordonnés correspondait assez bien avec la qualité des cibles articulatoires. Ainsi, Les personnes situées en bas du plan sont toutes des personnes dont l'articulation des syllabes et phonèmes est mauvaise. L'axe des abscisses en revanche correspondrait davantage à la fluence avec à gauche des personnes avec une fluence normale et à droite, celles avec une

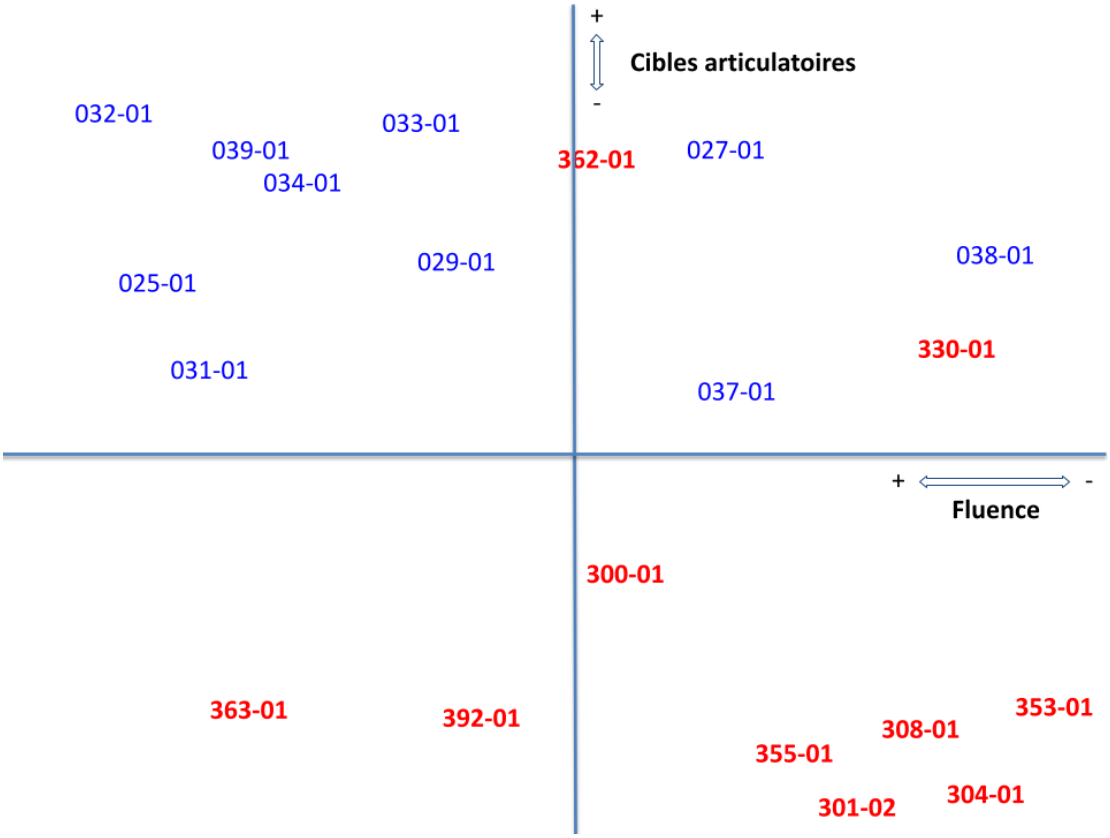

FIGURE 3.12 – Catégorisation libre d'une sélection de 10 patients cancer VADS (en rouges, en gras) et 10 sujets témoins (en bleus). Les numéros sont les identifiants des locuteurs, ils seront réutilisés dans les prochaines figures. Les axes (déterminés après la catégorisation) indiquent la fluence en abscisse (bonne fluence à gauche) et la qualité des cibles articulatoires en ordonnée (bonnes cibles articulatoires en haut)

fluence dégradée. Les personnes considérées comme ayant une mauvaise fluence sont principalement des locuteurs dont la segmentation de la parole est anormale avec par exemple la présence de pauses anormalement fréquentes. Nous retrouvons donc en haut à gauche, les locuteurs sains ne présentant pas de troubles de la parole à l'opposé des personnes dans la partie inférieure droite qui ont des troubles sévères avec une mauvaise fluence et de mauvaises cibles articulatoires. À l'interface entre ces deux extrêmes, nous retrouvons des regroupements entre des sujets témoins et pathologiques. La zone supérieure droite contient par exemple des personnes avec une bonne articulation mais une fluence dégradée. Nous retrouvons dans cette zone deux patients atteints de cancer VADS ainsi que trois sujets contrôles. Les troubles de la fluence chez les sujets contrôles (ou patients) peuvent alors correspondre à des problèmes de lecture qui entraînent des disfluences.

Grâce à cette catégorisation, nous avons donc une représentation nous permettant d'évaluer la proximité des caractéristiques rythmiques des locuteurs. De fait, lorsque

nous testerons les modélisations automatiques du rythme, nous pourrons nous assurer que deux personnes proches dans cette catégorisation possèdent des paramètres automatiques similaires.

3.2.4 Évaluation des variations prosodiques

En plus des annotations concernant la structuration hiérarchique des énoncés des locuteurs et la catégorisation libre, nous avons pu obtenir, grâce aux travaux de Anna Marczyk durant son post-doctorat sur le projet RUGBI, un ensemble de scores sur différents aspects précis de la prosodie des patients (Marczyk et al., en cours). L'objectif principal de ces travaux était de pouvoir déterminer sur quelles caractéristiques se base un annotateur devant évaluer la prosodie, la voix et l'articulation d'une personne. En effet, l'évaluation de ces trois dimensions de la parole incluent de nombreuses caractéristiques. Un premier évaluateur jugeant la prosodie d'un enregistrement pourrait par exemple se baser principalement sur les variations de f_0 de la voix alors qu'un second serait plus focalisé sur le rythme. Cette étude a été menée sur l'ensemble des enregistrements du corpus cancer (3.1.2).

Dans cette expérience, les enregistrements ont été découpés en plusieurs lots afin de réduire le nombre d'annotations par session pour les huit juges. Chaque lot a été constitué de sorte à avoir un équilibre des locuteurs au niveau de leur sévérité clinique perceptive. Dans chaque lot, tous les enregistrements de lectures de texte ont été évalués par huit juges naïfs. Pour chaque locuteur, il a été demandé aux juges de noter de façon globale le degré d'altération des trois dimensions : prosodie, voix et articulation. En plus de ces informations globales, un ensemble de 17 caractéristiques prosodiques et vocales ont été proposées, chacune appartenant à une dimension générale. Les 17 caractéristiques évaluées sont :

Dimension prosodique :
- Vitesse d'élocution
- Durée des syllabes
- Durée des mots
- Durée des pauses
- Placement des pauses
- Groupement des mots
- Variations mélodiques
- Variations rythmiques
- Monotonie
- Caractère naturel
- Accentuation

Dimension vocalique :
- Intensité
- Hauteur
- Voix éraillée
- Désonorisation
- Effort vocal
- Parole à bout de souffle

Les juges ont alors évalué chacun de ces 17 paramètres (+ les trois paramètres globaux de prosodie, voix et articulation) à la suite de deux écoutes maximum. L'évaluation a été faite sur une échelle à cinq niveaux : "Très anormal", "Clairement anormal", "Légèrement anormal", "Très légèrement anormal" et "Normal". Un des intérêts de cette étude provient du choix de faire participer des juges naïfs et non des experts

ORL. Cette particularité nous permet de pouvoir comparer les scores des experts avec ceux des juges naïfs sur les trois dimensions globales. Ainsi, nous pourrons savoir si les experts utilisent les mêmes critères que les naïfs sans être influencés par leur biais de connaissance de la parole pathologique. Parmi les 17 critères, les plus intéressants pour nos travaux sont ceux correspondant à la dimension prosodique de la parole. Notamment les caractéristiques en lien avec le rythme de la parole dont : l'altération de la hauteur, de l'intensité, de la durée syllabique, la vitesse d'élocution, le placement des pauses, le groupement des mots, les variations mélodiques, les variations rythmiques, le caractère naturel ou monotone de la parole, ainsi que l'accentuation. L'évaluation de ces critères nous permettra alors potentiellement d'apporter une meilleure interprétabilité des modélisations du rythme que nous mettrons en place. Cela pourra nous aider à extraire et comprendre des caractéristiques de nos modèles en comparant les résultats de nos extractions automatiques avec ces jugements perceptifs. De même, ces dimensions pourront nous aider à mieux comprendre les dimensions prosodiques utilisées dans la catégorisation libre décrite précédemment.

L'interprétation des résultats de cette expérience est à l'heure actuelle encore en cours et ceux-ci feront prochainement l'objet d'une publication (Marczyk, en cours). Afin de réaliser une analyse préliminaire des résultats, nous avons voulu comparer les différents critères entre eux ainsi qu'avec les scores perceptifs cliniques.

Afin de vérifier cela, nous avons calculé la matrice de corrélation entre les différents critères évalués ainsi que les scores de sévérité et d'intelligibilité des experts ORL. Cette matrice de corrélation est disponible dans la figure 3.13.

Notons tout d'abord que les scores de sévérité et intelligibilité cliniques ont été évalués de sorte à avoir une valeur proche de 10 pour les personnes avec une intelligibilité (ou sévérité) normale. À l'inverse, pour les autres scores, plus la valeur est élevée, plus l'altération est forte. Les indices de sévérité et intelligibilité sont donc inversement corrélés avec les autres paramètres seulement à cause de cette inversion d'échelle de notation. Nous pouvons observer sur la matrice que les paramètres sont globalement corrélés entre eux. Ceci est un résultat attendu étant donné que certains critères représentent des caractéristiques très proches les unes des autres. Nous retrouvons ainsi une forte corrélation entre la variabilité rythmique et la variabilité mélodique qui sont des notions liées. Le score clinique évaluant la prosodie des patients est le plus corrélé avec les indices de monotonie ($r = 0,66$) et de débit de parole ($r = 0,65$). Ainsi, il semblerait que les cliniciens jugent l'altération de la prosodie en se basant en grande partie sur ces dimensions. Néanmoins, il est intéressant de remarquer que l'évaluation clinique de la prosodie n'est corrélée qu'à 0,51 avec l'accentuation jugées par le panel d'auditeurs naïfs. De même, le placement et la durée des pauses ne semblent pas être pris en compte dans l'évaluation des scores experts. La construction des scores cliniques des médecins semble difficile à décomposer. Il est compliqué de savoir quelles caractéristiques de la paroles sont utilisées lorsqu'ils jugent la prosodie par exemple. Il sera donc nécessaire dans de futurs travaux de pousser davantage l'analyse de ces scores perceptifs de sorte à améliorer la compréhension que nous avons des jugements

perceptifs cliniques. Connaître précisément ces critères nous permettrait alors de réaliser des modélisations automatiques plus pertinentes de ces scores.

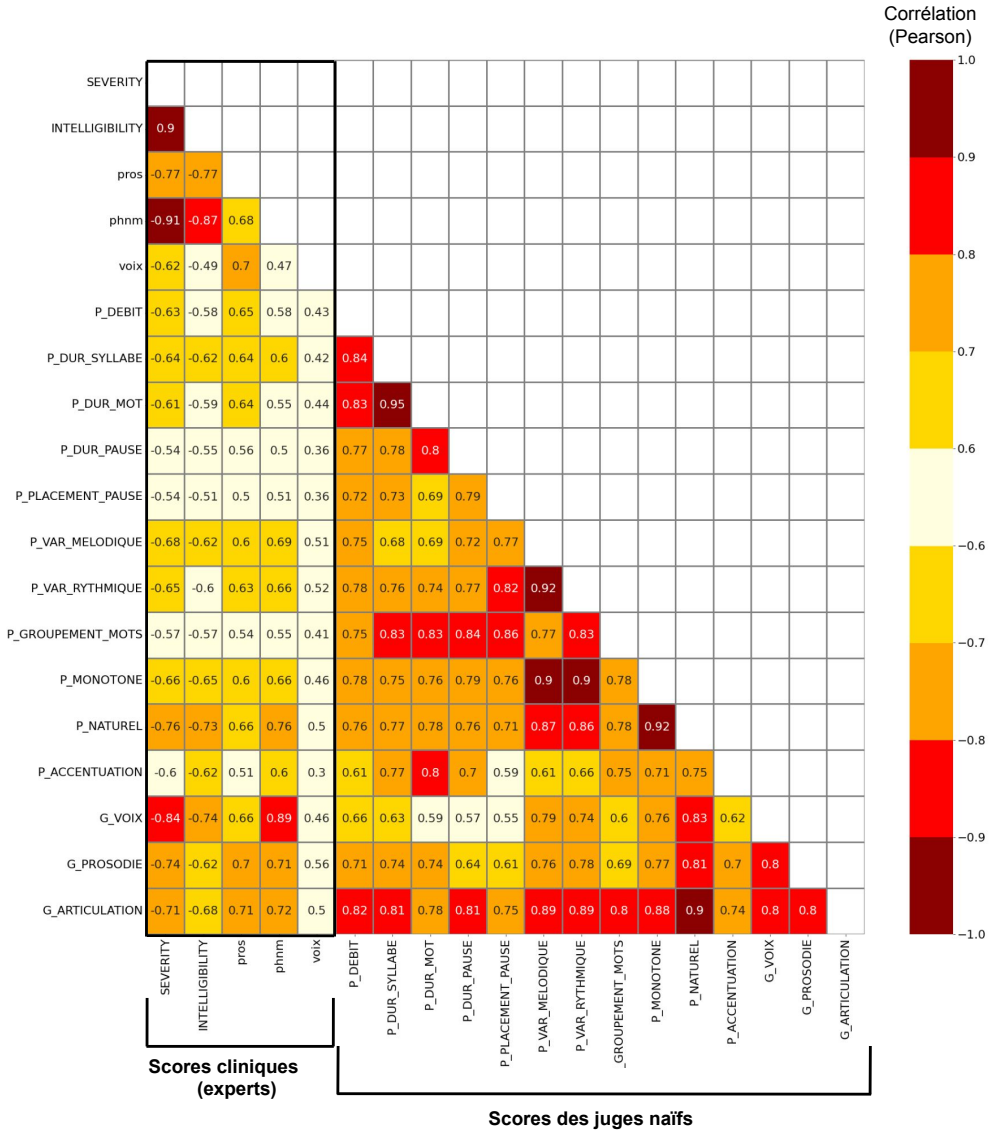

FIGURE 3.13 – Matrice de corrélation des différents indices sur l'ensemble des lots évalués par huit auditeurs naïfs. Les scores des huit juges ont été moyennés afin d'obtenir un score par enregistrement. Les cinq premières dimensions (sévérité, intelligibilité, prosodie, phonèmes, voix) sont celles issues du jugement perceptif clinique (section 3.1.5). Les 11 suivantes sont les caractéristiques de la dimension prosodique évaluées par le panel de huit auditeurs naïfs (P_...). Les trois dernières sont les caractéristiques globales de la parole (G_Voix ; G_Prosodie ; G_Articulation) également évaluées par les juges naïfs.

3.3 Conclusion de chapitre

Dans ce chapitre, nous avons décrit les corpus de parole continue que nous avions à notre disposition. Dans un premier temps, nous avons décrit un petit corpus de slam ayant été annoté au niveau rythmique (Simon, 2020). Ce corpus composé de parole fondamentalement rythmique nous permettra de tester la pertinence des modèles du rythme dans le chapitre suivant. Par la suite, nous avons présenté un corpus de parole de personnes atteintes de cancers des voies aéro-digestives supérieures. Ce corpus est composé de 113 locuteurs (87 patients et 26 témoins). Chaque locuteur a enregistré une dizaine de tâches de production de parole. Nous avons cependant noté un léger déséquilibre entre patients et témoins où environ une dizaine de sujets contrôles ne sont pas totalement appariés en âge. Le corpus Parkinson quant à lui est un grand corpus (205 patients et 111 sujets témoins) davantage équilibré mais qui présente également des limitations. Ces limitations proviennent principalement de la faible variété de troubles de la prosodie de la parole chez ces patients ce qui complexifie davantage nos travaux dont l'objectif est la caractérisation de ces troubles. Ces deux corpus proposent un ensemble de données cliniques avec notamment des scores d'évaluation de la sévérité et de l'intelligibilité des locuteurs estimés par des professionnels de la santé. Nous avons pu observer que le corpus cancer VADS possède une diversité plus large que le corpus Parkinson au niveau de ses scores. En plus de ces enregistrements et de leurs données cliniques, nous avons pu récupérer et créer un ensemble d'annotations spécialement pour l'évaluation de la prosodie des patients avec des annotations de la fréquence fondamentale dans le but de savoir quels algorithmes d'estimation de la f_0 sont les plus performants dans le cadre de ces pathologies. Les résultats de cette évaluation seront détaillés dans le chapitre suivant dans la section 4.4.1. De plus, nous avons annoté les différentes structurations prosodiques produites par un sous-ensemble du corpus cancer VADS sur la tâche de lecture de texte. Ces annotations nous aideront à interpréter au mieux nos modélisations automatiques du rythme en nous indiquant quels sont les niveaux prosodiques mis en jeux. Une catégorisation libre de ce même sous-ensemble a également été réalisée. Cette catégorisation a pour but de modéliser le rythme que nous percevons chez les patients en nous indiquant quelles personnes sont les plus proches les unes des autres. Cette information de proximité entre les locuteurs nous permettra plus tard d'évaluer la pertinence de nos paramètres automatiques du rythme en comparant ces paramètres chez des personnes proches dans la catégorisation. Enfin, nous avons pu récupérer un ensemble de jugements perceptifs dont l'objectif était de pouvoir identifier au mieux quels aspects de la parole sont évalués dans le cadre de la mesure de l'altération de la prosodie sur le corpus de parole cancer. Cette expérience nous a permis de déterminer plus précisément quelles dimensions de la prosodie sont les plus corrélées entre elles. Cela nous permettra alors de pouvoir extraire automatiquement des caractéristiques prosodiques pertinentes de nos modélisations.

Maintenant que nous avons décrit les différentes données que nous avons à notre

disposition, nous allons pouvoir sélectionner et mettre en place un ensemble de modélisations automatiques du rythme de la parole sur ces corpus de parole pathologique. Ces modélisations seront alors plus tard utilisées pour vérifier si l'étude du rythme de la parole peut permettre d'aider à l'amélioration de la mesure d'intelligibilité, et nous pourrons vérifier à l'aide de nos annotations prosodiques expertes et naïves si les modèles que nous allons mettre en place permettent ou non de caractériser efficacement le rythme de la parole pathologique.

4

Éprouver les modélisations du rythme sur la parole continue

Sommaire

Nous avons donc vu dans le chapitre 1 que le rythme de la parole est une structure hiérarchique qui organise l'accentuation à différents niveaux. Le chapitre 2 nous a montré les différentes modélisations automatiques du rythme qui existent à ce jour et enfin le chapitre 3 permet d'avoir un aperçu de la difficulté de travailler sur de la parole pathologique de par la difficulté d'obtenir des annotations prosodiques objectives et fiables. À partir de ces informations, nous avons dû nous focaliser sur les modélisations rythmiques qui respectent plusieurs conditions : premièrement la modélisation doit pouvoir rendre compte de *l'aspect hiérarchique* de la parole en observant des régularités à différents niveaux. Deuxièmement, il est préférable d'en choisir une qui permet de s'affranchir au maximum d'annotations extérieures. Enfin, si possible, l'étude du rythme doit pouvoir se faire en fonction de la durée du signal de parole,

c'est à dire que les caractéristiques rythmiques doivent pouvoir varier et être observées dans le *décours temporel* de la parole. Nous avons donc porté notre attention sur deux modélisations principalement qui sont le tempogramme (2.2.2) et les spectres de modulations d'amplitude (2.3.1) et de fréquences (2.3.2). Nous n'avons en revanche pas étudié les modulations spectro-temporelles (2.3.3), bien qu'elles remplissent certaines de nos conditions, car elles fournissent à notre sens une représentation difficile à interpréter à ce jour. L'interprétation des résultats se faisant majoritairement en isolant les dimensions (fréquentielle et temporelle), il nous semble plus pertinent de nous focaliser directement sur les modulations de fréquence et d'amplitude séparément afin d'analyser leurs contributions respectives. Nous dirigeons tout de même le lecteur vers l'étude de Marczyk et collab. (2022) qui a analysé les modulations spectro-temporelles sur notre corpus de parole cancer VADS.

4.1 Le tempogramme

Notre premier choix a été de nous tourner vers l'adaptation du tempogramme de Le Coz (2014) à l'analyse de la parole. L'algorithme du tempogramme (décrit dans la section 2.2.2) consiste à segmenter le signal en indiquant les frontières de zones stables au travers d'une segmentation forward-backward (F/B) (Andre-Obrecht, 1988) et de calculer une transformée de Fourier sur les pics pondérés de cette segmentation. Cependant, bien que la segmentation forward-backward soit utile pour détecter des zones stables, elle divise le signal de parole en segments de durées inférieures au phonème. Or, nous avons pu voir dans le chapitre 1 que l'unité optimale pour l'étude du rythme de la parole du français était la syllabe. Nous avons donc souhaité utiliser une segmentation différente, plus adaptée à l'étude de la parole, en nous basant sur une unité proche de la syllabe. Il est également intéressant de noter que nous n'avons pas utilisé la méthodologie du tempogramme proposée par Grosche et collab. (2010), car le calcul des accentuations du signal se base sur l'apparition de variations d'énergies soudaines dans le spectre, ce qui est pertinent dans le cadre de la recherche d'apparition de notes en musique, mais l'est beaucoup moins en parole, car cela marcherait sur la prononciation de consonnes occlusives comme /p/, /t/ ou /k/ mais ne ferait pas ressortir suffisamment les variations moins abruptes dans le signal des autres phonèmes.

4.1.1 Le Tempogramme appliqué à la parole

Dans un premier temps, nous avons voulu essayer d'observer comment se comporte cet algorithme sans aucune modification dans le cadre d'une parole très contrôlée afin de pouvoir interpréter au mieux les résultats générés. Nous avons alors essayé de calculer le tempogramme sur un signal très simple et régulier : une énumération de chiffres. Un exemple de tempogramme sur une énumération de 1 à 28 est disponible sur la figure 4.1.

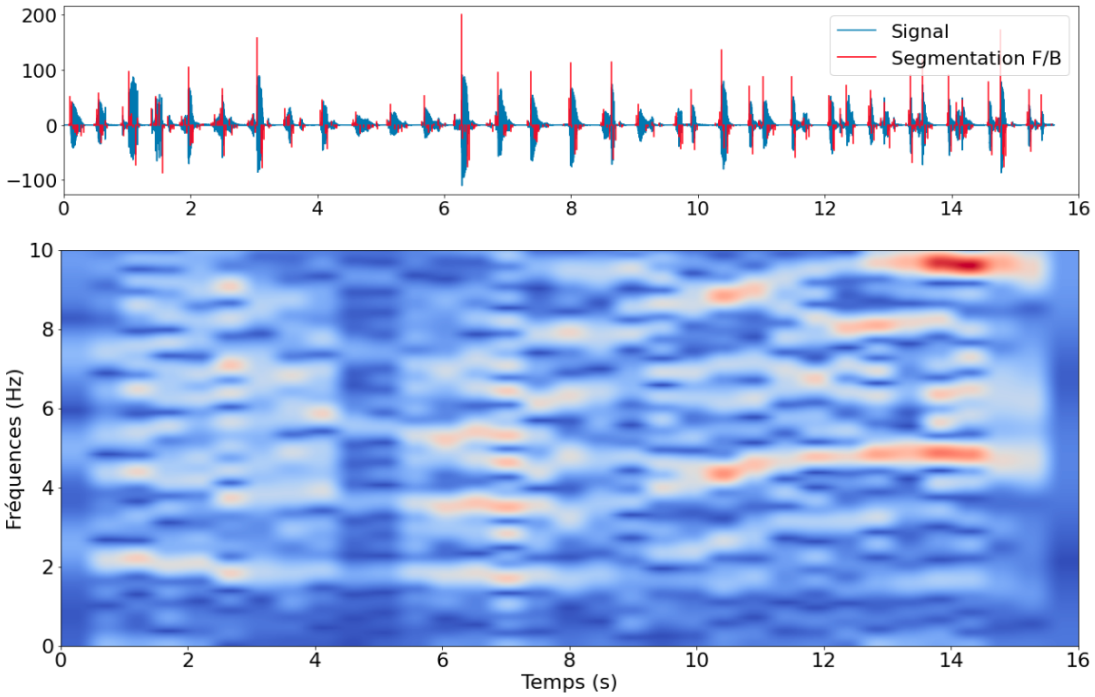

FIGURE 4.1 – Exemple de calcul de tempogramme sur un signal d'énumération (de 1 à 28). La partie haute représente le signal en bleu avec les segments F/B pondérés en rouges. En dessous, le tempogramme montre l'évolution dans le temps des différentes fréquences misent en jeu.

On peut alors observer que durant les huit premières secondes, une bande d'énergie apparaît aux alentours de 2 Hz, ce qui correspond au débit du comptage avec une prononciation d'environ deux nombres par seconde. À partir de dix secondes, cette régularité disparaît au profit d'une autre aux alentours de 4 Hz qui, cette fois, correspond aux nombres à partir de vingt. Cette nouvelle régularité s'explique par le fait que la majorité des nombres à partir de là sont bi-syllabiques alors que les précédents étaient davantage mono-syllabiques. La régularité principale détectée est donc celle des syllabes tandis que la régularité au niveau des mots est fortement atténuée.

À partir de cet exemple, il semble donc assez facile d'analyser et interpréter les résultats obtenus avec le tempogramme. Avant d'appliquer cette méthodologie à notre corpus de parole pathologique, il nous a semblé pertinent de la tester sur un corpus de parole fondamentalement régulière rythmiquement : le slam (3.1.1). L'analyse du tempogramme sur différents extraits de slam nous permettrait alors de vérifier s'il est possible d'observer les régularités de la parole à différents niveaux prosodiques.

Nous nous sommes alors particulièrement intéressés à deux extraits de slam (cf. tableau 3.1), le premier issu du titre "Slam" qui est un slam classique scandé (non chanté) particulièrement régulier dans ses durées entre deux syllabes accentuées (voir figure 3.1).
En effet, les durées inter-accentuelles sont généralement entre 500 et 800 ms et les syllabes accentuées sont également marquées fortement par un allongement de leur

durée avec des longueurs environ deux fois plus longues que les syllabes inaccentuées en moyenne. Le niveau prosodique dominant de cet extrait est le syntagme accentuel (ap). Nous espérons donc trouver une régularité correspondante à ce niveau dans la représentation en tempogramme.

Le second extrait (voir figure 4.2) est un slam chanté (sans instruments). Cet extrait est pour sa part marqué davantage par une forte régularité au niveau du mot prosodique : ils ont des durées équivalentes autour de 700 ms. Également, une importante régularité des durées de syllabes inaccentuées, autour de 200 ms, est observable.

Nous avons alors calculé les représentations en tempogramme de ces deux signaux en utilisant l'implémentation de Le Coz (2014). Les exemples de tempogrammes sont affichés dans les figures 4.3 et 4.4.

Le tempogramme de "Slam" (figure 4.3) semble exhiber une récurrence autour de 3, 4 et 5 Hz sur les quatre premières secondes du fichier, ce qui ne correspond à aucun des niveaux prosodiques que nous avons constatés comme étant réguliers. Concernant le tempogramme du fichier "Dandy" (figure 4.4), il montre principalement la régularité syllabique évoquée précédemment autour de 200 ms, qui se traduit par la présence importante d'une bande d'énergie à 5 Hz sur la figure. La régularité au niveau du mot prosodique n'est quant à elle pas retrouvée.

Nous avons donc pu observer que bien que le tempogramme fournisse une représentation visuelle très intuitive, il est difficile de tirer des interprétations claires de celle-ci. Le tempogramme rend principalement compte de la régularité syllabique et écrase les évènements périodiques qui pourraient se dérouler à des niveaux prosodiques supérieurs (inférieurs à 4 Hz).

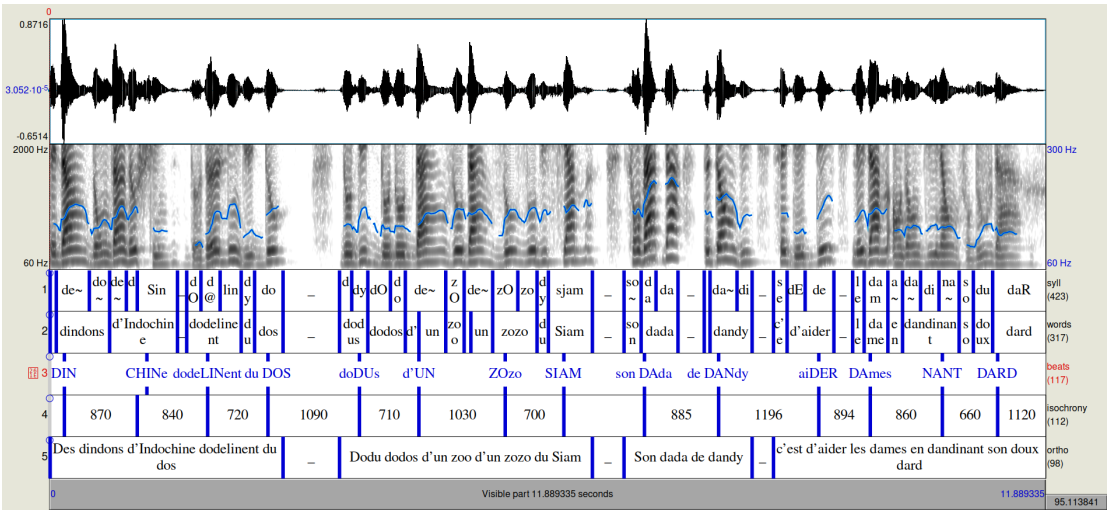

FIGURE 4.2 – Extrait du titre "Dandy" avec une analyse des durées inter-accentuelles. Analyses prosodiques réalisées par Anne Catherine Simon (Simon, 2020).

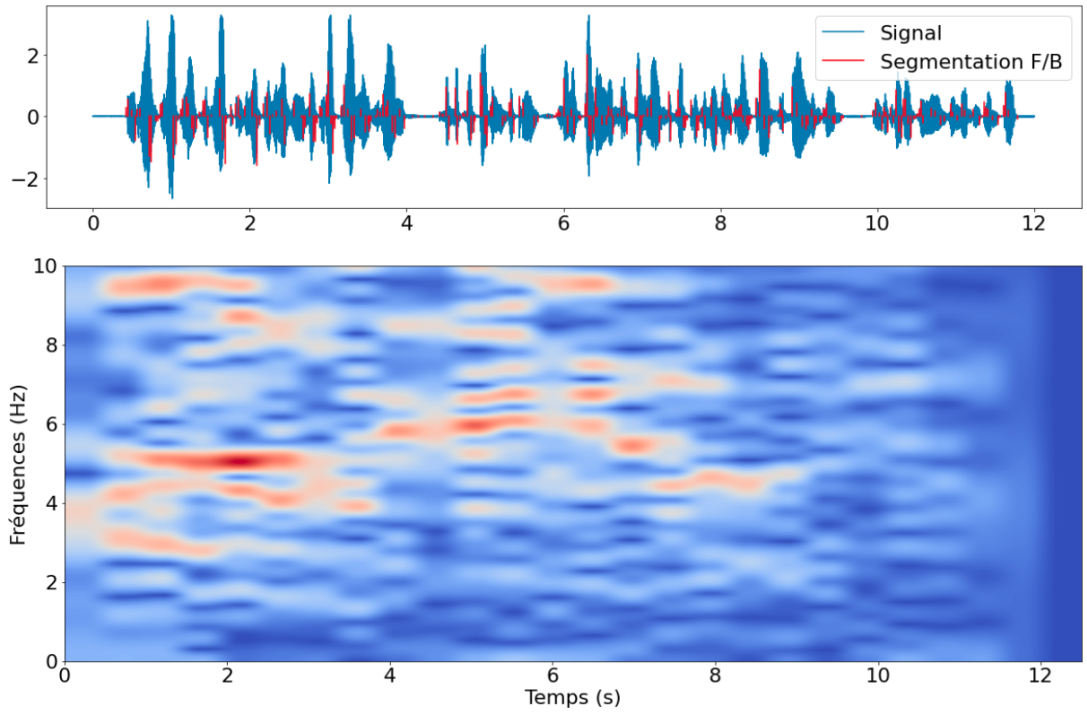

FIGURE 4.3 – Calcul du tempogramme sur l'extrait du titre "Slam".

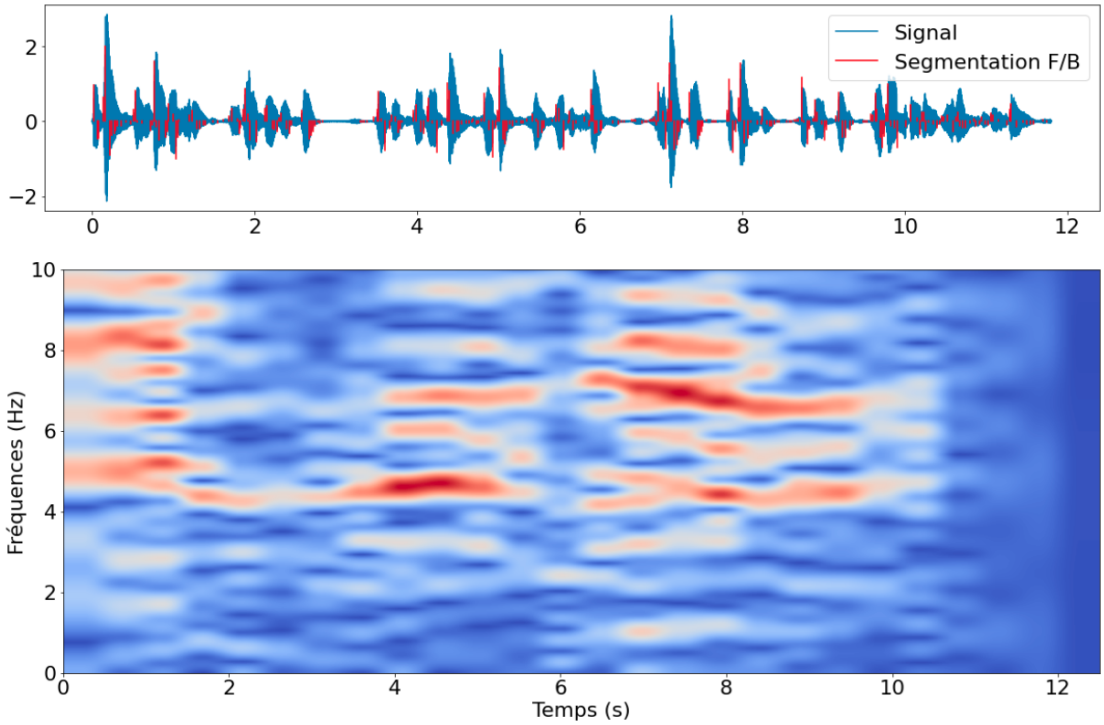

FIGURE 4.4 – Calcul du tempogramme sur l'extrait du titre "Dandy".

4.1.2 Adaptation de la segmentation

Comme nous l'avons vu dans la section 2.2.2, la segmentation *Forward / Backward* Andre-Obrecht (1988) est une segmentation très fine du signal à un niveau inférieur

au phonème. Cependant, dans le cadre de l'étude du rythme de la parole, nous avons vu que l'unité prosodique minimale était la syllabe. Nous avons donc pensé à remplacer la segmentation *Forward / Backward* par simplement une segmentation en syllabes afin de limiter les interférences avec des unités plus petites.

Les frontières utilisées par le tempogramme ont alors été placées au niveau du début des voyelles de chaque syllabe. De plus, nous avons également conservé la pondération par la différence d'énergie après et avant chaque frontière. Un exemple de résultat obtenu via une segmentation syllabique est disponible sur la figure 4.5.

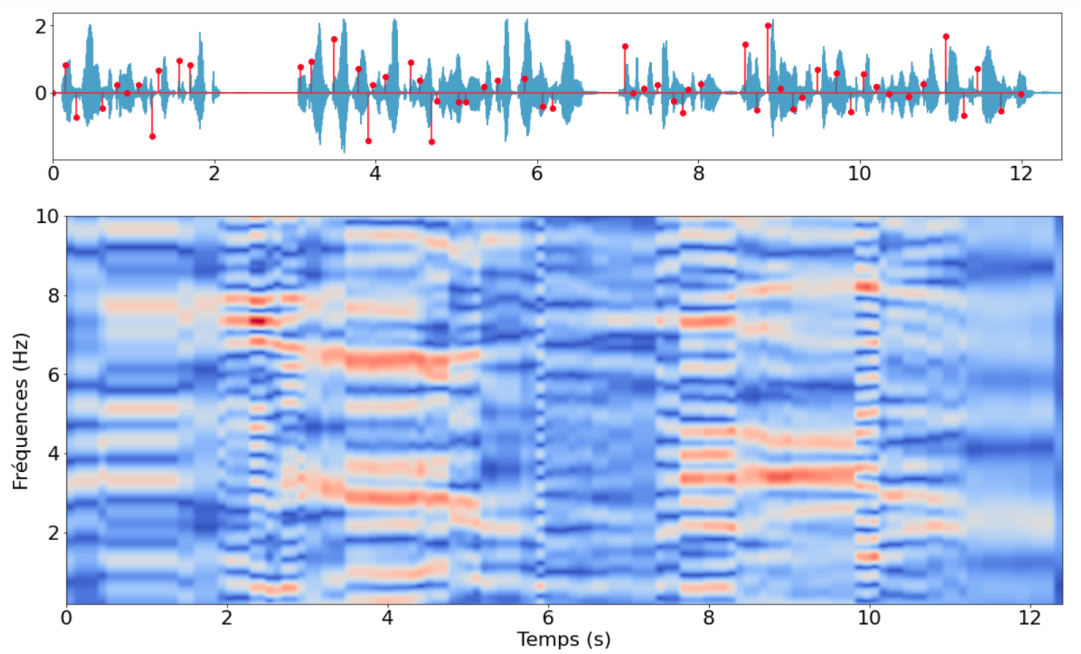

FIGURE 4.5 – Calcul du tempogramme sur l'extrait du titre "Slam" en utilisant une segmentation en syllabes. La segmentation superposée au signal brut est visible en haut et le tempogramme qui en découle est en dessous.

Nous observons alors que le tempogramme ainsi obtenu produit un grand nombre d'informations sur des régularités qui apparaîtraient à différents niveaux. Cependant, nous pensons que seules quelques bandes de fréquences sont pertinentes, et qu'une grande majorité ne serait que le résultat d'une somme d'harmoniques. C'est à dire que dans le cas où une fréquence autour de deux Hertz est présente, cela implique également que son amplitude se répercutera à des fréquences multiples (4, 6, 8, 10...) . Ces multiples sont appelées des harmoniques et sont par exemple visibles sur l'exemple du comptage (figure 4.1) où la bande de fréquence de 2 Hz aux alentours de 1 seconde se répercute plus haut dans les bandes 4, 5 et 8 Hz. Ce phénomène courant dans l'utilisation de la Transformée de Fourier est d'autant plus marqué par l'utilisation de pics abrupts pour matérialiser la segmentation. En effet, la segmentation produit un signal où l'ensemble des valeurs sont à zéro à l'exception des frontières où une seule valeur est modifiée. On parle alors de pics de la distribution de Dirac (Dirac (1930) p. 60). Lorsqu'une transformée de Fourier est appliquée à un signal périodique de

fréquence ω composé uniquement de pics de Dirac, alors les amplitudes du spectre seront toutes égales aux fréquences multiples de ω. Afin d'éviter ce phénomène, nous avons voulu essayer de remplacer les pics de Dirac par des pics moins abrupts en les remplaçants par des courbes gaussiennes. Cela permet de réduire l'intensité des harmoniques. Un exemple de signaux composés de pics de Dirac et de gaussiennes et leurs transformées de Fourier est visible sur la figure 4.6.

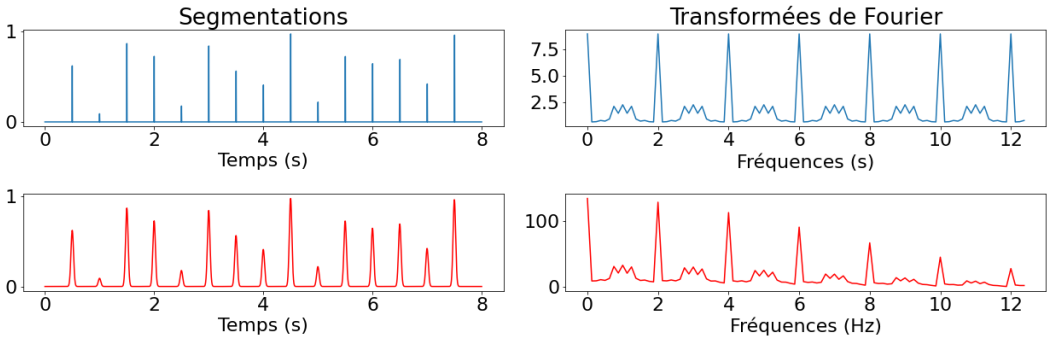

FIGURE 4.6 – Exemple de transformées de Fourier (à droite) sur des signaux (à gauche) composés de pics de Dirac (en bleu) et de de courbes gaussiennes (en rouge)

Les valeurs des harmoniques sont donc atténuées progressivement en utilisant des courbes gaussiennes tandis que sur les pics de Dirac tous les harmoniques ont des amplitudes égales. En appliquant ce processus à nos signaux audio, on obtient le signal suivant :

$$S = \sum_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(t-s_i)^2}{2\sigma^2}} \tag{4.1}$$

Avec n le nombre de syllabes, s_i l'onset de la i-ème syllabe, t un vecteur du temps et σ l'écart type des gaussiennes. Nous avons empiriquement choisi un écart-type de 20 millisecondes. Un exemple de tempogramme obtenu avec cette méthode est disponible sur la figure 4.7.

Par rapport au tempogramme sur la figure 4.5, quelques bandes fréquentielles "parasites" ont été atténuées, mais une majorité d'entre elles sont malheureusement toujours visibles et rendent l'interprétation du tempogramme difficile. Malgré l'amélioration de la visualisation, un désavantage de cette méthode réside dans le coût de calcul du tempogramme. En effet, dans le cas des pics de Dirac, le calcul de la transformée de Fourier est très simple étant donnée que la grande majorité du signal est mis à zéro. Lors du passage aux courbes gaussiennes, ce nombre diminue et le calcul est donc davantage coûteux en temps bien qu'il reste tout de même relativement rapide à exécuter. Ainsi, le calcul du tempogramme sur un fichier audio de 30 secondes se fait en moyenne en 1,6 secondes pour la version avec les pics contre 3,1 secondes avec les gaussiennes (test réalisé sur un ordinateur portable équipé d'un i7-8550U).

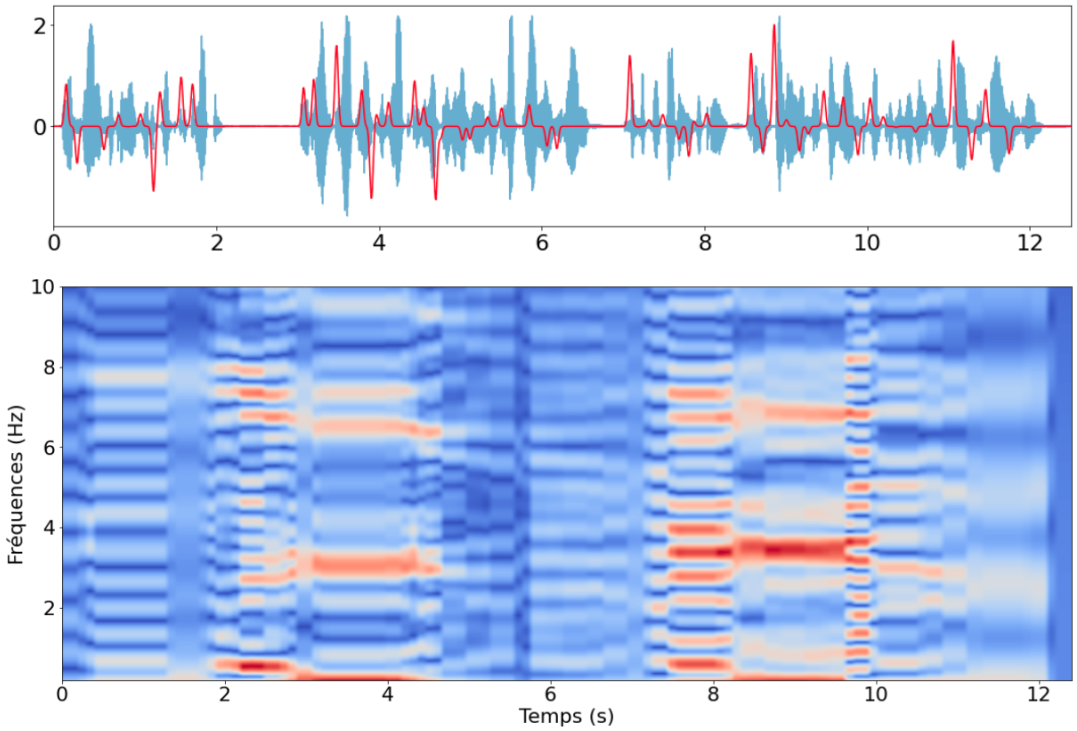

FIGURE 4.7 – Calcul du tempogramme sur l'extrait du titre "Slam" en utilisant une segmentation en syllabes et le remplacement des pics par des courbes gaussiennes. La segmentation superposée au signal brut est visible en haut et le tempogramme qui en découle est en dessous.

4.1.3 Développement d'un plugin Praat

Afin de rendre l'extraction et la visualisation du tempogramme plus simple, nous avons développé une extension au logiciel Praat (Boersma, 2001) permettant d'extraire le tempogramme d'un signal audio. Ce logiciel est un outil très utilisé au sein de la communauté des linguistes ; c'est pourquoi nous avons choisi d'essayer de porter l'implémentation du tempogramme sur celui-ci.

Le développement d'extension Praat étant relativement limité pour créer de nouvelles visualisations, nous avons décidé d'utiliser l'objet Spectrogramme déjà existant au sein du logiciel afin de générer facilement des tempogrammes. En effet, le tempogramme et le spectrogramme ont de nombreuses caractéristiques communes ; il est donc tout à fait pertinent d'utiliser cet objet pour simplifier le processus. Un premier exemple de tempogramme extrait sous Praat est disponible sur la figure 4.8.

Comme nous pouvons le constater, sous Praat les valeurs sont obligatoirement en nuances de gris et le contraste affiché pour le tempogramme est très faible. Cela ne permet donc pas d'analyser correctement les différentes fréquences prépondérantes dans le signal. Nous avons donc décidé, afin d'accentuer le contraste, de modifier les valeurs du tempogramme en leur appliquant une fonction sigmoïde définie comme suit :

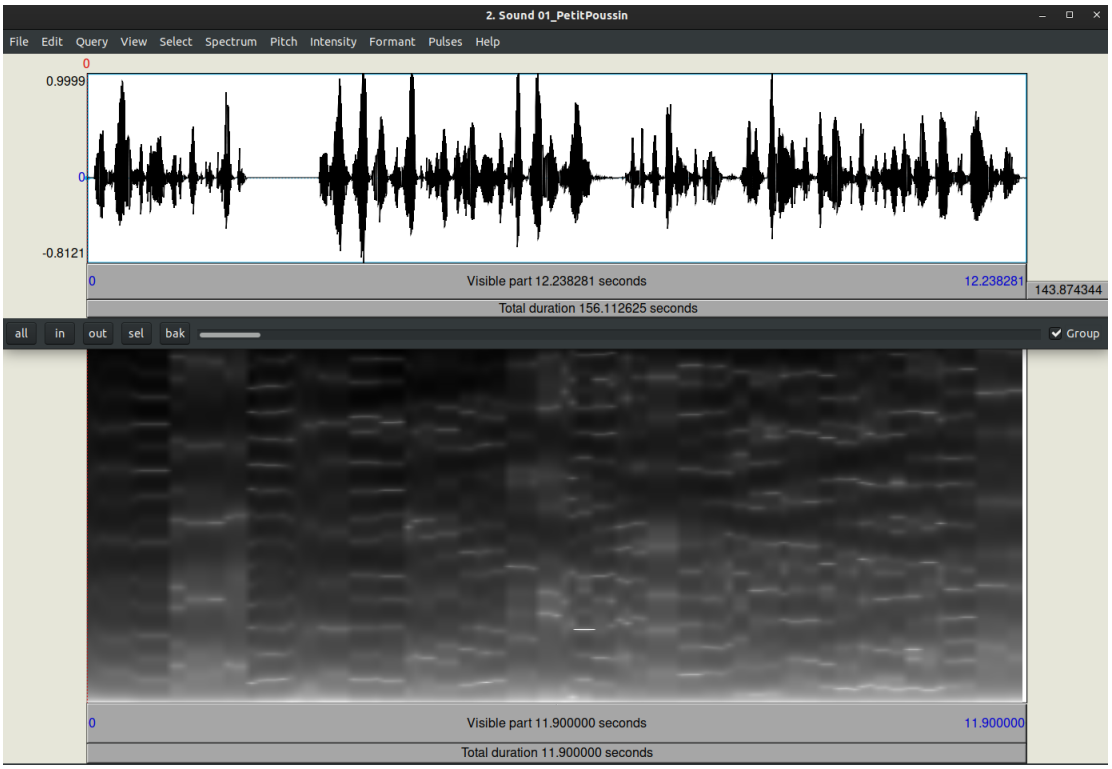

FIGURE 4.8 – Calcul du tempogramme via l'intégration au logiciel Praat sur l'extrait du titre "Slam".

$$T'_{ij} = \frac{1}{1 + e^{-\lambda T_{ij}}}$$

Où T_{ij} représente la valeur du tempogramme normalisée entre -1 et 1 à la ligne i et la colonne j. La fonction sigmoïde permet de limiter les effets de contraste dans le cas où quelques valeurs isolées du tempogramme sont très fortes (ou très faibles), ces valeurs seront ramenées à 1 (ou -1). Cette normalisation permet alors d'obtenir des résultats beaucoup plus visibles comme le montre la figure 4.9.

La limitation principale de cette normalisation est qu'il devient alors difficile de distinguer quelles sont les fréquences les plus dominantes parmi les couleurs les plus foncées. D'autres méthodologies auraient pu être employées afin de pallier cela (comme peut-être une égalisation d'histogramme), mais nous n'avons pas souhaité améliorer davantage cette représentation afin de nous consacrer à l'étude d'autres représentations du rythme.

4.1.4 Limites du tempogramme

Comme nous avons pu le voir, le tempogramme est un outil avec un potentiel intéressant pour étudier les régularités rythmiques de la parole à différents niveaux. Dans les expériences que nous avons menées, nous n'avons cependant pas réussi à obtenir des résultats fiables quant à l'analyse d'une régularité rythmique à un niveau autre

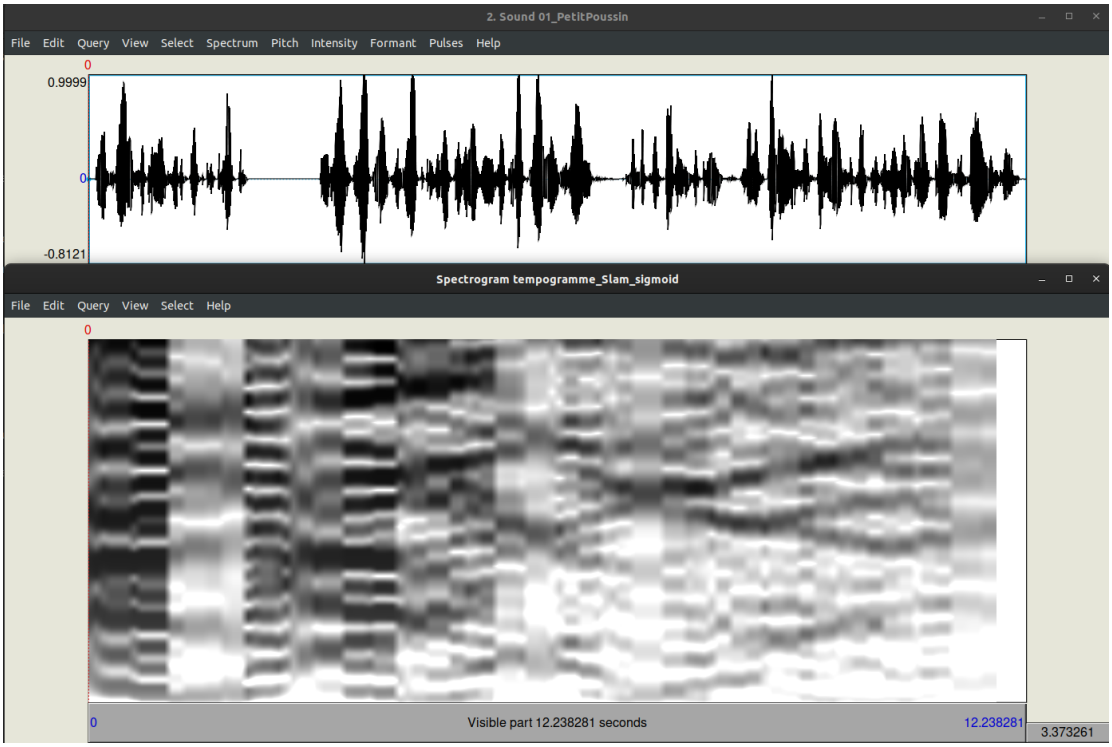

FIGURE 4.9 – Calcul du tempogramme via l'intégration au logiciel Praat sur l'extrait du titre "Slam". Les valeurs du tempogramme sont normalisées par une fonction sigmoïde.

que celui de la syllabe. En effet, en analysant des textes de slam dont les régularités rythmiques sur des unités plus grandes que la syllabe sont évidentes, nous ne sommes pas parvenus à les observer. De plus une problématique majeure de cette modélisation est la forte présence de fréquences "parasites" qui apparaissent à des valeurs harmoniques des fréquences dominantes. La présence de ces harmoniques dans toutes les bandes de fréquences implique donc des difficultés pour localiser précisément les régularités du signal. Comme nous l'avons vu (4.1.2), le passage de pics à des courbes gaussiennes permet d'améliorer légèrement les résultats, mais pas suffisamment pour pouvoir exploiter cette modélisation.

Nous pensons donc que le défaut majeur du tempogramme provient de l'utilisation d'une segmentation en amont de la transformée de Fourier. Cela implique une perte conséquente de l'information contenue initialement dans le signal. Nous pensons donc qu'il est nécessaire d'utiliser une modélisation du rythme qui ne se base pas sur une segmentation préalable du signal. Ainsi, nous conserverons une quantité d'information suffisante pour pouvoir interpréter au mieux les résultats des transformées de Fourier. C'est le cas par exemple des modélisations basées sur les modulations du signal que nous avons évoquées dans la section 2.3.

4.2 Le Spectre de Modulations d'Amplitude

Afin de pallier les limites du tempogramme décrites ci-dessus, notre choix de modélisation s'est donc tourné vers le spectre d'enveloppe d'amplitude (que nous noterons EMS pour *Envelope Modulation Spectrum*) décrit dans la section 2.3.1. Pour rappel, ce modèle consiste à extraire l'enveloppe d'amplitude du signal et d'appliquer une transformée de Fourier à cette dernière. Le point fort de cette modélisation, par rapport au tempogramme, est le matériau de base fourni à la transformée de Fourier. En effet, tandis que le tempogramme fournit une succession de pics qui ne représente pas toujours très bien les modulations rythmiques du signal, l'enveloppe d'amplitude quant à elle permet de décrire l'évolution dans le temps du signal de parole.

4.2.1 Mise en œuvre

Bien que nous voulions utiliser une méthodologie semblable à celle de Tilsen et Johnson (2008), nous nous sommes interrogé quant à la pertinence de l'extraction l'amplitude d'enveloppe. Pour rappel, dans leur étude, ils effectuent un filtre passe-bande entre 700 et 1300 Hz sur le signal suivi d'un passage en valeur absolue, d'une soustraction de la moyenne et enfin d'un filtre passe-bas à 10 Hz afin de ne conserver que l'information des modulations lentes du signal. Nous avons donc ré-implémenté cette méthode en python afin de tester la pertinence de cette enveloppe sur notre corpus. Une illustration de l'extraction d'enveloppe est présentée sur la figure 4.10.

Bien que l'enveloppe ainsi extraite semble convenir pour modéliser les variations lentes de modulations du signal, nous pouvons également remarquer que le premier filtrage entre 700 et 1300 Hz proposé initialement par Cummins et Port (1998a) a tendance à accentuer certains sons. C'est le cas notamment des voyelles /a/ dont l'énergie augmente fortement par rapport aux autres. De plus le son /y/ est fortement atténué alors que dans le signal initial, le mot "eu" est le plus énergétique. Nous pensons que cette modification d'intensité dans l'enveloppe est principalement due au fait que le filtrage entre 700 et 1300 n'est pas adapté aux voyelles du français. En effet, certaines voyelles comme le /y/ ou le /i/ ont leurs deux premiers formants en dehors de la plage 700-1300 Hz tandis que la voyelle /a/ (F1 = 700 Hz, F2 = 1200 Hz) possède ses deux formants dans cette bande. De fait, nous obtenons des amplifications ou des diminutions de l'intensité en fonction de la place des voyelles dans le triangle vocalique.

Pour compenser cela, nous avons donc testé de nombreux intervalles de filtrage, et celui que nous avons retenu est un premier filtre à 300-1000 Hz. Ainsi, nous incluons l'ensemble des premiers formants des voyelles françaises en excluant la majorité des seconds formants. Un exemple de comparaison entre le filtrage initial et le nôtre est visible sur la figure 4.11.

Nous pouvons alors voir que les amplitudes des différents phonèmes sont davantage respectées dans notre calcul de l'enveloppe. Par exemple, les /a/ de "n'avait" et

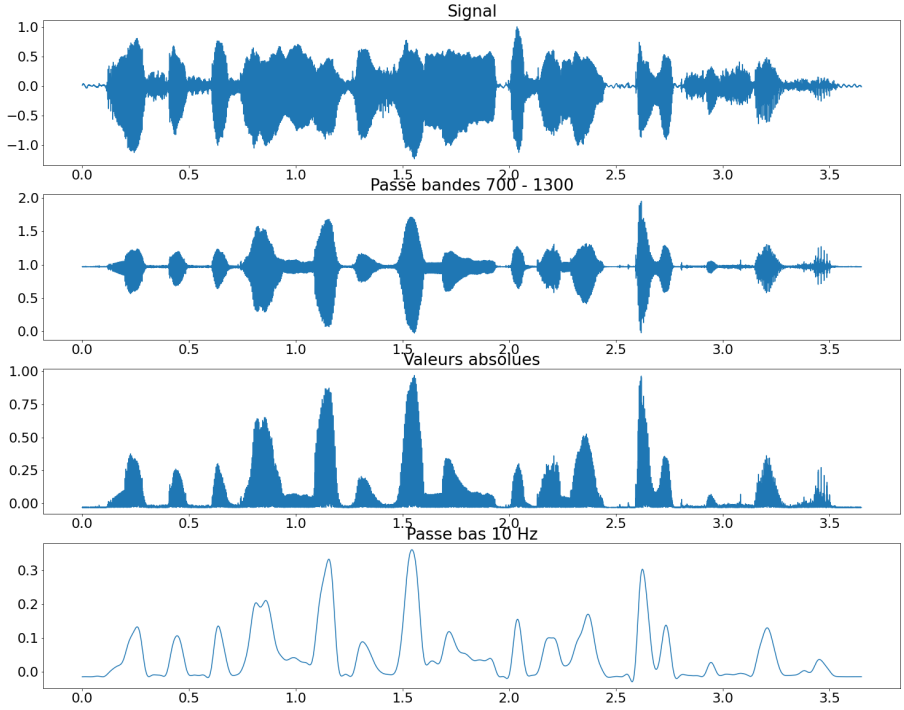

FIGURE 4.10 – Illustration du processus de calcul de l'enveloppe d'amplitude sur un signal de parole saine ("Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres").

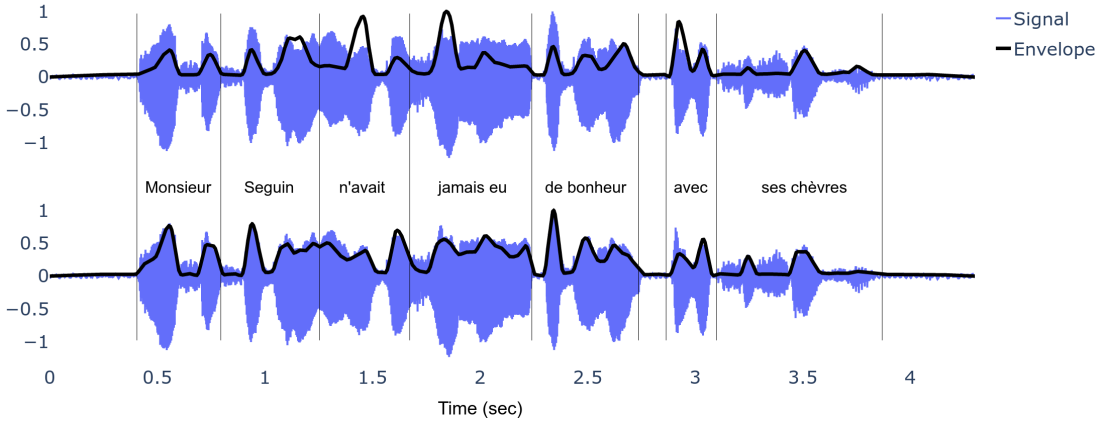

FIGURE 4.11 – Comparaison du calcul de l'enveloppe d'amplitude. En haut, le premier filtre passe bandes est effectué entre 700 et 1300 Hz, en bas, le filtre est entre 300 et 1000 Hz

"jamais" ne sont plus exagérés. En revanche, les consonnes nasales /n/ et /m/ dont le

premier formant se situe entre 300 et 700 Hz deviennent alors plus énergétiques dans notre méthode alors qu'elles étaient effacées avec le filtrage à 700-1300 Hz. Malgré ce léger défaut, nous avons décidé de conserver notre filtrage qui selon nous est plus polyvalent.

4.2.2 Transformée de Fourier de l'enveloppe

Une fois que l'enveloppe de modulation a été extraite, il reste à étudier les régularités présentes dans l'enveloppe. Cette étape est réalisée via l'utilisation d'une transformée de Fourier tout comme dans le cas du tempogramme. Cela permet alors de déceler des formes (ou motifs) qui se répètent au sein de l'enveloppe de modulation d'amplitude. Ces formes peuvent par exemple être des syllabes (accentuées ou non) ou des combinaisons de syllabes (mots prosodiques, groupes accentuels...) qui apparaissent à de multiples reprises. Plus un motif de durée t se répète de façon régulière tout au long de l'enveloppe, plus la valeur du spectre correspondant à la fréquence $\frac{1}{t}$ sera élevée. Prenons par exemple un EMS d'une portion du fichier "Dandy" visible en figure 4.12.

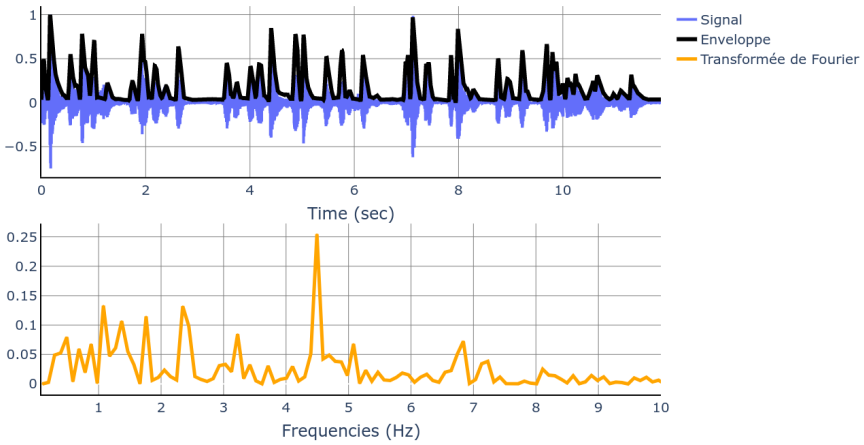

FIGURE 4.12 – Transformée de Fourier d'une enveloppe de modulation d'amplitude appliquée à l'extrait du slam "Dandy" décrit dans la figure 4.2. En haut, l'enveloppe de modulation en noir est superposée au signal brut en bleu. En bas, la Transformée de Fourier de l'enveloppe pour des fréquences inférieures à 10 Hz.

Pour rappel, une analyse prosodique de cet extrait est visible sur la figure 4.4. Nous avions alors vu que cet extrait présente une très forte régularité syllabique aux alentours de 200 ms ainsi qu'une plus légère régularité inter-accentuelle aux alentours de 700 ms. La régularité syllabique apparaît très clairement sur l'EMS avec un pic vers 4,5 Hz (222 ms) qui est la plus haute valeur du spectre. En revanche la régularité inter-accentuelle est moins présente, mais elle reste tout de même visible avec un ensemble de pics entre 1 et 2,3 Hz (soit 1 s et 430 ms). Sur ce fichier de parole chantée, l'EMS semble donc fournir des résultats pertinents par rapport aux régularités prosodiques que nous avions pu observer manuellement.

Nous pouvons alors nous demander si cela fonctionne de la même façon sur l'extrait "Slam" qui, lui, est scandé. Le résultat sur cet extrait est donné en figure 4.13. Sur ce

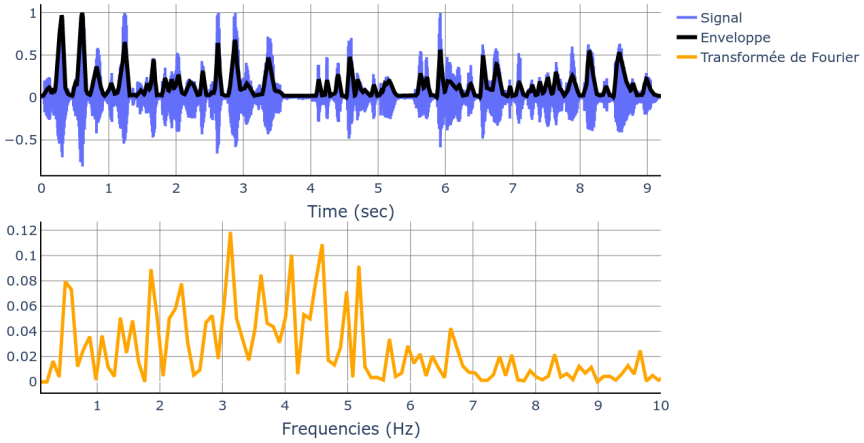

FIGURE 4.13 – Transformée de Fourier d'une enveloppe de modulation d'amplitude appliquée à l'extrait du slam "Slam" décrit dans la figure 3.1. En haut, l'enveloppe de modulation en noir est superposée au signal brut en bleu. En bas, la Transformée de Fourier de l'enveloppe pour des fréquences inférieures à 10 Hz.

passage, les résultats sont moins évidents que sur "Dandy", mais il est clair que de nombreuses régularités apparaissent dans des unités inférieures à 6 Hz (supérieures à 160 ms). Les deux plus hauts pics se situent à 3,13 Hz (320 ms) et 4,6 Hz (217 ms). Le premier pourrait correspondre à la récurrence des mots prosodiques sur la première phrase (*Une grosse droite ou un coup d'latte qui vous retourne et vous éclate l'âme, le slam*) dont les durées avoisinent les 320 ms ("une grosse", "droite", "un coup"...). Le deuxième pic quant à lui pourrait représenter l'alternance de certaines longues syllabes sur la dernière phrase (*Sans naphtaline et sans formol, c'est bien plus grisant que l'alcool, ça vole le slam*).

L'EMS réussit donc à nous fournir des informations quant aux informations prosodiques contenues dans un énoncé. Cependant, il peut parfois nous fournir un trop grand nombre d'information qui peuvent cacher les informations principales. Il apparaît alors judicieux de supprimer les informations inutiles et de ne ressortir que les zones de fréquences principalement misent en jeu dans la production orale. Pour cela, nous pourrions par exemple utiliser un lissage du spectre afin d'agréger les informations contenues dans des pics adjacents. Cela permet alors d'avoir une information plus globale quant à la structure de l'énoncé.

4.2.3 Lissage de l'EMS

La présentation de la courbe EMS présente de nombreuses variations locales. Afin de faciliter la lecture, et de pouvoir mieux localiser les zones les plus énergétiques, nous avons procédé à un lissage de cette courbe.

Nous avons réduit le nombre de points du spectre pour n'avoir environ que trois points tous les Hz. Cette réduction a été effectuée en appliquant une moyenne glissante toutes les 0,15 Hz. La courbe finale est ensuite obtenue en interpolant par un polynôme de degré 3 (Akima (1970)). Cette courbe passe par tous les points moyens. La figure 4.14 illustre sur un exemple l'extraction du spectre (en orange) et la résultante lissée (en pointillé rouge).

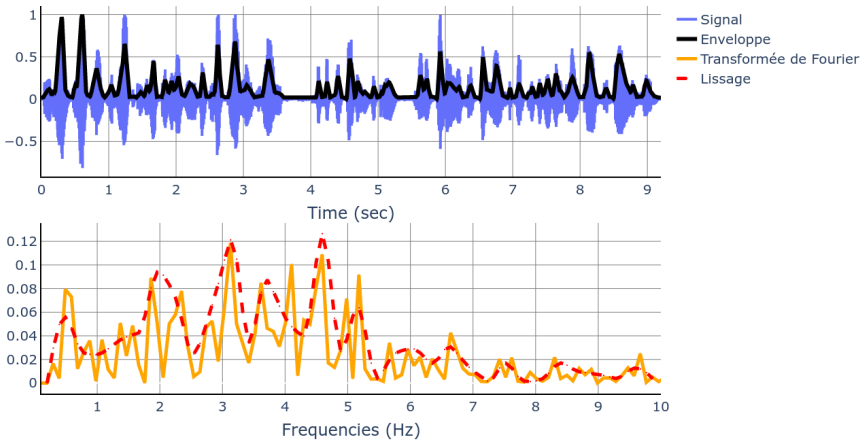

FIGURE 4.14 – Exemple de lissage de l'EMS sur un extrait du texte "Slam". L'EMS est en orange et le lissage résultant est un rouge.

Cette représentation lissée permet de détacher du spectre les zones les plus énergétiques. Notre objectif était de rendre plus lisible les agrégats énergétiques autour des valeurs courantes des IP, ip, ap, pw, syllabe et phonèmes (cf. section 1.1.3 pour la description des niveaux prosodiques et section 1.2.3 pour les considérations sur leur durée) : soit pour les IP et ip entre 1 et 2 s ($0,5 - 1Hz$), entre 450 ms et 600 ms pour ap ou pw ($1,6 - 2,2Hz$), environ 250 ms pour les syllabes ($4Hz$) et environ 100-150 ms pour les phonèmes ($6 - 10Hz$).

4.3 Le Spectrogramme du rythme

Le spectre de modulation d'amplitude permet donc de localiser les zones de régularités rythmiques à différents niveaux dans le cadre d'un énoncé. Nous pensons qu'il pourrait être pertinent d'analyser les variations temporelles de ces fréquences et donc de générer un spectrogramme du rythme. Ce spectrogramme peut se calculer en extrayant l'EMS sur une fenêtre glissante de plusieurs secondes de la même manière que pour le tempogramme (4.1). Cette méthodologie a été réalisée en parallèle de nos travaux par Gibbon (2021) qui calcule un spectrogramme à partir de spectres d'enveloppe.

Notre calcul de spectrogramme est basé sur l'extraction d'un EMS sur une fenêtre

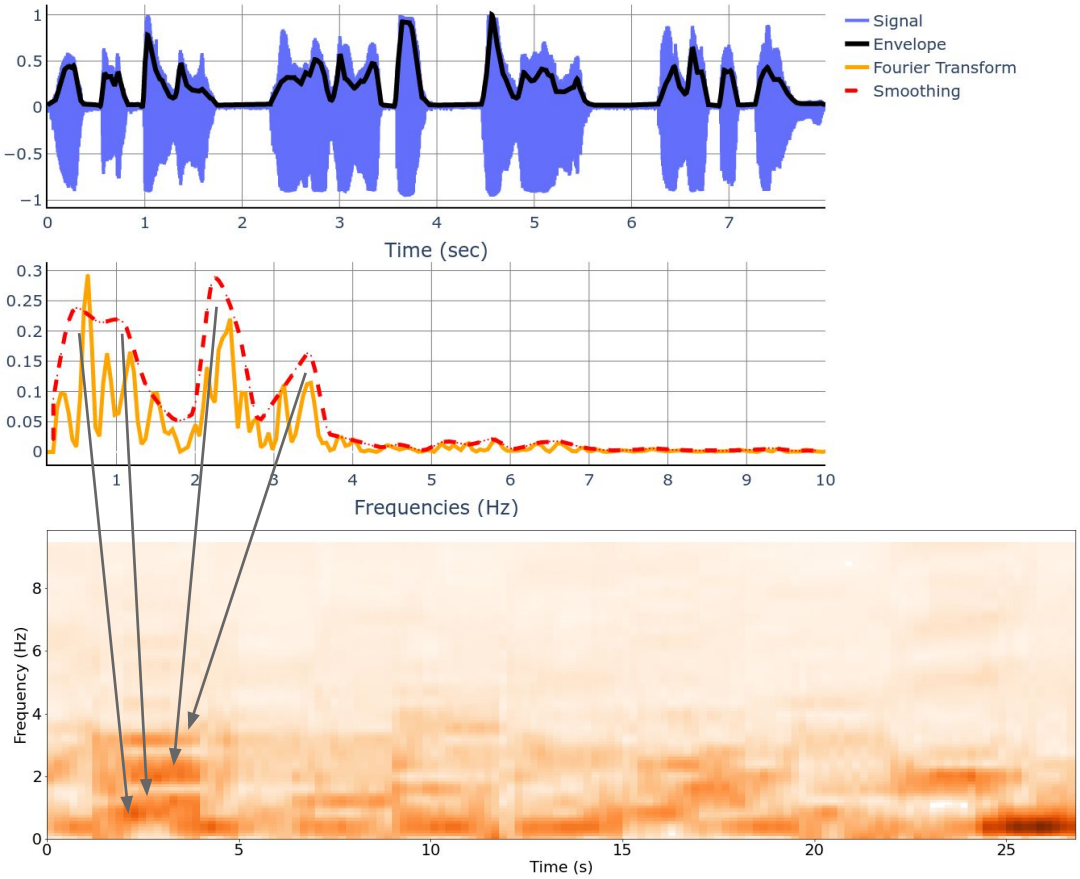

FIGURE 4.15 – Exemple de spectrogramme du rythme obtenu à partir d'une lecture de texte d'un locuteur sain (sujet n^o 1). La partie haute représente le signal et son enveloppe d'amplitude sur la phrase "Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres". En dessous, on retrouve le spectre de modulation d'amplitude et tout en bas le spectrogramme sur l'ensemble de la lecture. Les flèches montrent comment les pics du spectre apparaissent dans le spectrogramme.

glissante de 3 secondes avec un décalage temporel de 0,2 secondes. Ainsi, nous pouvons obtenir un spectrogramme comme illustré sur la figure 4.15.

Sur ce spectrogramme tiré d'une lecture du texte de la chèvre de Monsieur Seguin, nous pouvons voir sur les 5 premières secondes les régularités que nous retrouvons dans l'EMS de l'énoncé. Ainsi, nous avons les trois zones d'intérêt marquées dans l'EMS qui se retrouvent sur le spectrogramme. En revanche, il reste tout de même difficile de fournir des interprétations visuelles simples de cette modélisation. Il est notamment rare de voir apparaître des régularités de manière continue tout au long d'un discours de plusieurs phrases. Nous voyons alors plusieurs explications possibles. Premièrement, il se pourrait que notre méthode de calcul basée sur une fenêtre glissante ne soit pas adaptée. En effet, lorsque le signal contient des pauses, il est difficile pour notre algorithme d'extraire de l'information pertinente sur une fenêtre de trois secondes. Il est également possible que la régularité rythmique d'une personne ne soit simplement pas la même d'une phrase à une autre et qu'elle s'adapte en fonction

de divers paramètres comme la taille de la phrase, le contexte ou son interlocuteur (voir la notion d'oscillateurs couplés évoquée dans la section 1.2.2). Il est donc plus difficile de mesurer le rythme de façon continue et il serait plus pertinent de se concentrer sur les régularités rythmiques que l'on peut trouver au sein d'un seul énoncé. C'est la raison pour laquelle nous nous concentrerons désormais sur l'étude de phrases isolées, de sorte à ne pas être influencé par ces phénomènes.

4.4 Le Spectre des Modulations de Fréquence

Maintenant que nous avons pu voir qu'il était possible d'étudier le rythme en nous basant sur les variations d'amplitudes du signal, il nous semble pertinent de réaliser la même expérience en nous basant sur les variations de fréquence. Pour cela, il existe diverses méthodes comme nous l'avons vu dans la section 2.3.2. Dans cette thèse, nous avons seulement eu le temps de traiter l'étude des variations de f_0. Nous pensons en effet que la fréquence fondamentale joue un rôle majeur dans la perception du rythme. Il est donc intéressant de savoir si l'étude de ces variations ajoute de l'information complémentaire à celle fournie par les modulations d'amplitude.

4.4.1 Choix de l'algorithme d'estimation de la F0

Avant de pouvoir extraire le spectre de la fréquence fondamentale, nous nous sommes posé la question de savoir si les algorithmes d'estimation de la f_0 utilisés habituellement sur de la parole saine étaient performants dans le cadre de la parole pathologique. En effet, afin d'extraire une enveloppe cohérente, il est crucial d'avoir une bonne estimation de base de la fréquence fondamentale. L'objectif ici est de tester un ensemble d'algorithmes de détection de f_0 sur des enregistrements de personnes atteintes de cancers VADS ainsi que de la Maladie de Parkinson (MP) afin de savoir lesquels sont les plus adaptés. Pour cela, nous avons sélectionné 12 algorithmes de détection de f_0 en nous basant en partie sur une récente étude (Jouvet et Laprie, 2017) ayant comparé ces algorithmes dans la parole bruitée. Nous avons également ajouté plusieurs algorithmes récents basés sur des réseaux de neurones profonds. Il est possible de catégoriser ces différents algorithmes en trois catégories principalement :

— Les algorithmes temporels qui se basent sur des méthodes comme l'autocorrélation du signal afin de calculer une estimation de la f_0
— Les algorithmes fréquentiels qui transforment le signal dans le domaine fréquentiel (au travers d'un Transformée de Fourier) afin de déterminer quelle est la fréquence dominante dans le signal.
— Enfin, nous classons à part les algorithmes basés sur des réseaux de neurones profonds en considérant que nous ne savons pas exactement par quelle représentation du signal ils passent pour estimer les valeurs de f_0

En plus de ces méthodes, nous avons également souhaité intégrer deux techniques de combinaison d'algorithmes. La première est un vote médian entre cinq algorithmes

(cf. table 4.1). La seconde est une combinaison entre l'algorithme REAPER qui est efficace pour détecter le voisement et FCN-F0 qui fournit des estimations très proches des valeurs de références mais qui n'excelle pas dans la détection de voisement. La combinaison des résultats de ces deux algorithmes en prenant les valeurs de f_0 de FCN-F0 sur les zones voisées indiquées par REAPER propose de bonnes performances.

TABLE 4.1 – Liste des algorithmes testés et l'implémentation utilisé correspondante. Les trois dernières colonnes indiquent si l'algorithme est basé sur le domaine temporel ou fréquentiel du signal, ou bien s'il utilise des méthodes d'apprentissage profond (RNN). Les algorithmes comportant une "*" sont ceux qui ont été utilisés pour le vote médian.

Algorithme	Implémentation	Temporel	Spectral	RNN
ACF (Boersma, 2000)	Praat	X		
AMDF* (Ross et collab., 1974)	Snack Sound Toolkit	X		
REAPER* (Google-Open-Source, 2015)	github.com/google/REAPER	X		
RAPT (Talkin et Kleijn, 1995)	Snack Sound Toolkit	X		
Enhanced RAPT (Ghahremani et collab., 2014)	Kaldi	X		
Yin (de Cheveigne et Kawahara, 2002)	github.com/patriceguyot/Yin	X		
NDF* (Kawahara et collab., 2005)	STRAIGHT	X	X	
YAAPT* (Kasi et Zahorian, 2002)	MATLAB implementation	X	X	
SWIPE (Camacho et Harris, 2008)	Speech Signal Processing Toolkit		X	
PEFAC (Gonzalez et Brookes, 2014)	VOICEBOX		X	
CREPE (Kim et collab., 2018)	github.com/marl/crepe			X
FCN-F0* (Ardaillon et Roebel, 2019)	github.com/ardaillon/FCN-f0			X

L'évaluation de ces algorithmes a été réalisée sur le sous-corpus de 24 fichiers (8 sains, 8 VADS et 8 MP) décrit dans la section 3.2.1. Chaque algorithme a été évalué selon sa capacité à déterminer si une zone de parole est voisée ou non ainsi que selon sa capacité à calculer une estimation proche de la F0 de référence en utilisant les métriques suivantes :

— *Voicing Detection Errors* (VDE) qui décrit le pourcentage de valeurs de f_0 où l'algorithme considère qu'une valeur est voisée alors qu'en réalité elle ne l'est pas et inversement.

— *Gross Pitch Errors* (GPE) qui décrit le pourcentage de valeurs de f_0 où l'algorithme a produit une valeur éloignée de plus de 20% de la valeur de référence. *F0 Frame Errors* (FFE) qui décrit le pourcentage de valeurs de f_0 où l'algorithme s'est trompé tant au niveau du voisement que de la valeur. Cette métrique est une combinaison des deux précédentes.

Les résultats bruts obtenus sont décrits dans le tableau en annexe A. Les algorithmes se basant sur le domaine temporel du signal proposent de bons résultats sur la détection de voisement : ACF (Boersma, 2000), AMDF (Ross et collab., 1974) et REAPER (Google-Open-Source, 2015) (score aux alentours de 5% d'erreurs que ce soit pour la parole pathologique ou saine). Concernant la précision des estimations de la f_0, ce sont les algorithmes basés sur des réseaux neuronaux qui procurent les meilleurs résultats avec environ 1% d'erreurs grossières sur la parole cancer pour FCN-F0 (Ardaillon et Roebel, 2019) et moins de 0,5% sur la parole saine et Parkinsonienne. Le combinaison de REAPER et FNC-F0 nous semble être le meilleur compromis entre détection de voisement et estimation de la f_0. En revanche, la complexité algorithmique pour calculer cette combinaison n'est pas idéale étant donné qu'il est nécessaire d'exécuter ces deux méthodes. Il serait possible afin d'optimiser le processus d'isoler la détection de voisement effectuée dans REAPER afin de pouvoir la combiner avec FCN-F0 sans

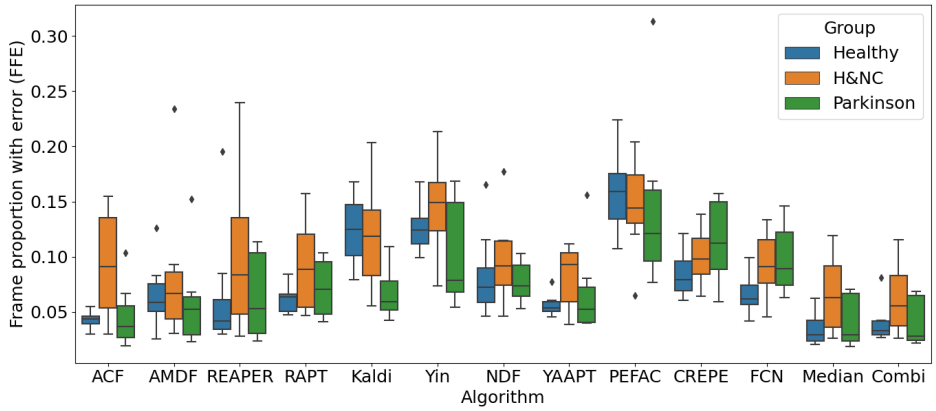

FIGURE 4.16 – Pourcentages d'erreurs de détection de f_0 en fonction du type de parole (saine en bleu, VADS en orange ou MP en vert) et de l'algorithme choisi. Figure extraite de Vaysse et collab. (2022a, p.3098)

réaliser de calculs superflus. Nous n'avons cependant pas encore exploré cette piste par manque de temps.

Les résultats détaillés de notre étude ont été publiés dans le journal international The Acoustic Society of America (Vaysse et collab., 2022a).

4.4.2 Le spectre de modulations de fréquence appliqué à la parole

Après avoir déterminé que l'utilisation d'une combinaison d'algorithme de détection de f_0 était une bonne solution pour extraire une courbe de f_0 fiable, nous avons pu effectuer des tests basés sur le calcul du spectre de la fréquence fondamentale. Pour cela, il existe à notre connaissance deux méthodologies. La première consiste à extraire la f_0 brute et d'y appliquer un *Root Lomb Periodogram* directement (Varnet et collab., 2017). Cet algorithme est en quelque sorte similaire à une Transformée de Fourier mais il est possible de l'utiliser sur des signaux discontinus. C'est le cas notamment de la fréquence fondamentale qui possède de nombreuses portions vides étant donné que la parole n'est pas tout le temps voisée. La seconde option est de considérer les zones non-voisées comme ayant une fréquence fondamentale égale à la médiane de la f_0. C'est le cas notamment de (Gibbon, 2021) qui considère ces zones comme étant pertinentes. De notre côté, nous pensons que les transitions abruptes de f_0 entre des zones voisées et non-voisées ne sont sont pas propices à être modélisées par une transformée de Fourier. C'est la raison pour laquelle nous avons choisi d'utiliser une interpolation des points de f_0 afin d'en déduire une courbe lissée. De cette manière, nous pourrons combler les parties non-voisées grâce à l'interpolation et donc appliquer la Transformée de Fourier.

Afin de modéliser la courbe intonative, nous avons choisi d'utiliser une interpolation cubique des points de f_0. Cela signifie que nous calculons entre chaque points un polynôme de degré trois dans le but de relier les points de façon naturelle. Un exemple

d'interpolation polynomiale est visible sur la figure 4.17. Cette méthodologie nous

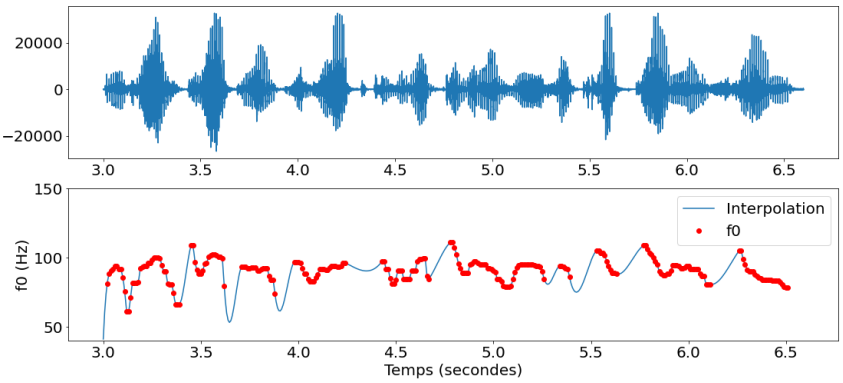

FIGURE 4.17 – Exemple de lissage de la f_0 par une interpolation polynomiale sur la phrase "*Une grosse droite ou un coup d'latte qui vous retourne et vous éclate l'âme, le slam*". Le signal de base est en haut, ses valeurs estimées de f_0 par l'algorithme de combinaison est en rouge, la courbe bleu est la fonction spline composée des différents polynômes.

fourni donc une représentation intéressante de la même forme que notre enveloppe d'amplitude (4.2). Nous avons donc calculé la transformée de Fourier de cette nouvelle enveloppe afin de voir si elle fourni des informations complémentaires à celle obtenue à partir de l'amplitude. Un résultat sur un extrait de slam est visible sur la figure 4.18

En comparant les deux courbes obtenues, nous voyons que les deux spectres sont similaires au niveau des fréquences misent en avant. Les deux pics principaux obtenus avec la modulation d'amplitude sont situés à 3 Hz (333 ms) et 3,75 Hz (267 ms). De même pour le spectre de modulations de fréquence, les deux pics principaux sont à 2,73 Hz (366 ms) et 3,61 Hz (277 ms). Les régularités misent en avant par ces deux modélisations sont donc extrêmement similaires. Dans cette forme, l'étude des modulations de fréquences ne présente donc que peu d'intérêt. Selon nous, cette limitation pourrait provenir de notre interpolation qui suit de façon trop stricte les valeurs de f_0. Cette contrainte impliquerait donc de suivre des variations micro-prosodiques non-pertinentes. La micro-prosodie concerne l'ensemble des variations de f_0 que nous pouvons détecter sur signal mais qui ne sont pas perçues par les auditeurs. Par exemple, à la suite d'une prononciation de consonne occlusive, "l'explosion" liée au relâchement de la pression sous-glottique peut entraîner une augmentation locale de la f_0. Ces phénomènes sont fortement présents dans ce texte de slam de par la présence de plusieurs occlusives, mais aussi par une accentuation initiale des mots prosodiques forte dans ce texte. La prise en compte de cette micro-prosodie va alors mettre davantage en avant les régularités basées sur les syllabes et les mots prosodiques. Cependant, un intérêt majeur selon nous de l'étude la f_0 est de pouvoir nous focaliser davantage sur la courbe intonative. Cette courbe intonative marque le domaine de l'unité intonative (IP) comme évoqué dans la section 1.1.3. Cette courbe intonative basée sur la f_0 n'inclut cependant pas les variations micro-prosodiques. Nous avons donc voulu essayer de

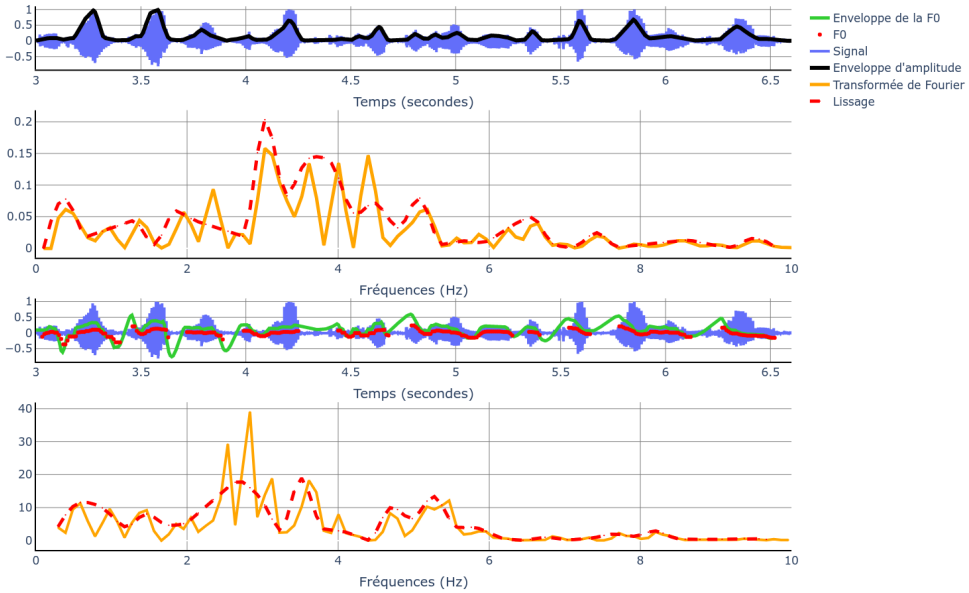

FIGURE 4.18 – Exemple de spectre de modulation de f_0 comparé au spectre de modulations d'amplitude. La moité haute concerne le spectre de modulations d'amplitude (en orange) et son lissage (en rouge). La moitié basse correspond au spectre de modulation de fréquence (en orange) appliqué à notre interpolation (en vert).

faire le même processus de lissage de la courbe de f_0 mais en utilisant cette fois une modélisation de l'intonation. Pour cela, nous avons utilisé l'algorithme Momel (Hirst et Espesser, 1993) qui modélise l'intonation au travers d'une approximation de la f_0 par une fonction spline qui est composée d'une succession de polynômes. À la différence de notre première approche, la nouvelle courbe ne passe pas par l'ensemble des points de la f_0. Elle se contente de la réduire à un ensemble de points caractéristiques. Ces points sont alors reliés entre eux par des polynômes de degré deux. La dérivée de la courbe aux points de jonction entre deux polynômes est déterminée de sorte à être nulle. Ces points de jonctions sont des points d'inflexions qui permettent alors d'obtenir une modélisation de la courbe intonative. Un exemple d'extraction de f_0 superposée à la courbe générée par Momel est visible sur la figure 4.19. . Sur cette figure, la courbe spline (en bleue) représente relativement bien les variations globales de f_0 tout en éliminant les pics correspondants à des phénomènes micro-prosodiques.

Une fois la courbe intonative modélisée, il suffit alors d'y appliquer une transformée de Fourier afin d'obtenir un nouveau spectre de modulation de la fréquence fondamentale. En reprenant le même exemple que sur la figure 4.18, nous obtenons le spectre visible sur la figure 4.20. Ce nouveau spectre de modulation de fréquences présente de fortes similitudes avec le précédent visible sur la figure 4.18. En effet, ces deux spectres possèdent des pics aux alentours de 2,7 Hz et 3,7 Hz. Ce nouveau spectre basé sur Momel permet cependant de mettre en avant les régularités de plus hauts niveaux avec un pic à 0,6 Hz (1,7 sec) qui émerge. Cela montre donc que cette méthode permet

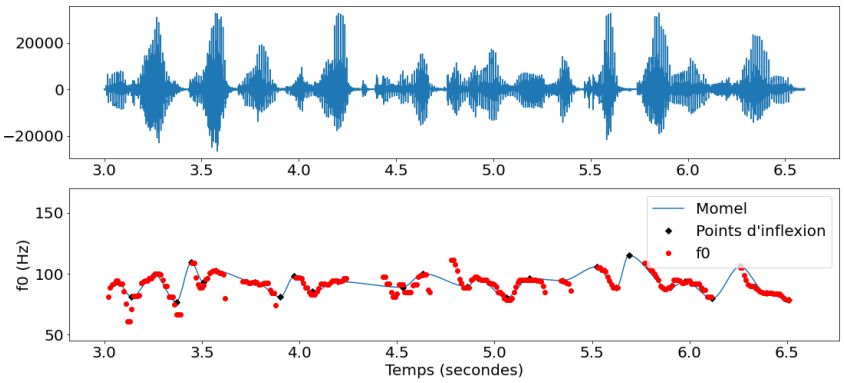

FIGURE 4.19 – Exemple de modélisation de la courbe intonative par l'algorithme Momel sur la phrase "*Sans naphtaline et sans formol, c'est bien plus grisant que l'alcool, ça vole le slam*". Le signal de base est en haut, ses valeurs estimées de f_0 par l'algorithme de combinaison est en rouge, la courbe bleu est la fonction spline créée par Momel et les points noirs sont les points d'inflexion.

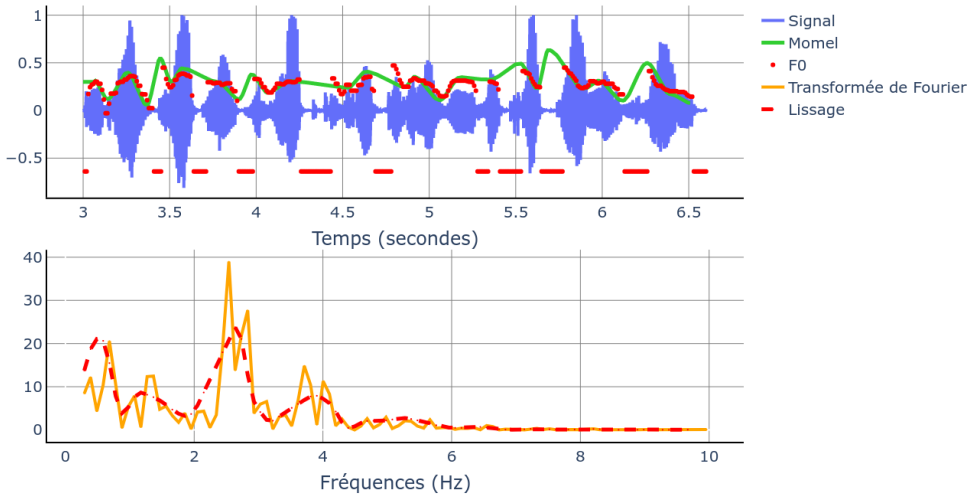

FIGURE 4.20 – Exemple de spectre de modulation de f_0. Le signal brut (en bleu) et sa courbe intonative (en vert) générée par Momel sont en haut. Le spectre de la courbe intonative est en bas. Le spectre brut est en orange et le spectre lissé est en rouge (en pointillés).

bien de mettre en avant les régularités des unités rythmiques les plus grandes. Or, nous avons vu dans le premier chapitre 1 que l'unité prosodique la plus large (l'unité intonative ou IP) était le domaine de la courbe intonative. Il est donc normal que cette modélisation mette davantage en avant ce type de régularité. De plus, les alternances rapides de syllabes ont tendance à être écrasées par Momel ce qui entraîne une forte diminution de l'énergie contenue dans la zone fréquentielle liée aux syllabes (> 4 Hz). De

même, les niveaux du pw ou de l'ap (< 4 Hz), domaines prosodiques minimaux inclus dans les unités prosodiques de rang supérieurs (typiquement, les unités intonatives ip ou IP), qui comportent des syllabes métriquement fortes pas forcément marquées par des variations de f_0, peuvent également être écrasés par la modélisation Momel. Il conviendrait donc, dans le futur, de combiner les spectres de modulation d'amplitude et de f_0 (via Momel) pour obtenir une visualisation complète du rythme de la parole.

4.5 Conclusion

Dans ce chapitre, nous avons vu comment nous avons implémenté et adapté les modélisations automatiques du rythme que sont le tempogramme et les spectres de modulations d'amplitude et de fréquence. Dans un premier temps, nous nous sommes intéressés au tempogramme qui est une méthode basée sur la transformée de Fourier d'une segmentation du signal. Cette méthode initialement créée pour étudier le tempo dans le cadre de la musique a dû être adaptée afin de correspondre au mieux à ce que l'on attend d'une modélisation du rythme. C'est pourquoi nous avons remplacé la segmentation qui était une segmentation de bas niveau par une segmentation des onsets syllabiques. En appliquant le tempogramme à des fichiers audios de slam, nous pensions pouvoir tirer facilement des interprétations de cette modélisation grâce à la structure rythmique forte de ce type de fichier. Cependant, malgré de nombreuses tentatives d'amélioration de l'algorithme, et le développement d'un plugin Praat pour faciliter nos expériences, nous ne sommes pas parvenus à interpréter ces résultats de façon reproductible. C'est la raison pour laquelle nous avons décidé de changer de modélisation pour nous concentrer sur les spectres de modulation d'amplitude et de fréquence. L'avantage de ces méthodes par rapport au tempogramme est qu'ils n'utilisent pas de segmentation du signal. Bien que le procédé soit le même que pour le tempogramme, la segmentation est remplacée par une extraction automatique des modulations d'amplitude ou de fréquence du signal. Ces modulations sont davantage informatives et permettent de conserver l'ensemble des variations lentes du signal initial.

Nous avons vu que le spectre de modulations d'amplitude (EMS) permet de modéliser efficacement la structure prosodique d'un énoncé de slam en nous permettant de voir les différents niveaux de régularité mis en jeu par le locuteur. Afin d'avoir une représentation plus simple et lisible, nous avons réalisé un lissage du spectre dans le but de supprimer les informations superflues et d'agréger les informations similaires en regroupant des pics proches. Ce lissage nous permet alors de localiser rapidement les niveaux prosodiques les plus réguliers dans un signal de parole. Enfin, nous avons voulu réaliser cette modélisation en utilisant des modulations de la fréquence fondamentale en lieu et place de l'amplitude. Utiliser la f_0 implique évidemment de la calculer. Pour cela, de nombreux algorithmes existent et produisent d'excellents résultats sur la parole saine. Or, dans la suite de cette thèse, nous allons travailler sur des fichiers de parole pathologique. Cependant, les algorithmes de f_0 n'ont pas été

développés dans ce but et certains ne produisent pas de résultats convaincants. Nous avons donc évalué ces algorithmes sur un sous corpus de parole pathologique afin de connaître le ou les algorithmes les plus pertinents. Nous avons alors pu voir qu'un algorithme combinant deux méthodes, une basée sur de l'autocorrélation du signal comme REAPER (Google-Open-Source, 2015) et une autre basée sur une approche de réseau de neurones comme FCN-F0 (Ardaillon et Roebel, 2019) est un bon compromis entre performance et vitesse d'exécution. Une fois l'algorithme choisi, nous avons pu comparer deux méthodologies visant à rendre la courbe de f_0 continue. La première consiste à calculer une interpolation polynomiale passant par tous les points de la f_0. La seconde se base sur une approximation de la courbe intonative appelée Momel (Hirst et Espesser, 1993). Ces deux méthodes proposent des modélisations fiables du rythme une fois la transformée de Fourier appliquée à ces enveloppes. Cependant, celles-ci génèrent des résultats très similaires au spectre de modulation d'amplitude. La seule différence que nous avons pu observer se trouvait sur le spectre de modulations utilisant Momel. Le spectre obtenu permettait alors de mettre davantage en avant les régularités rythmiques d'unités prosodiques larges comme les ip ou les IP. Cependant ce gain de précision dans les unités prosodiques larges est compensé par une forte perte en précision dans la régularité des fréquences supérieures à quatre Hertz (domaine des syllabes et phonèmes).

Nous avons donc vu différentes méthodes afin d'extraire des enveloppes utilisées pour le calcul du spectre de modulation. À la suite de nos expériences, nous pensons que l'utilisation des modulations d'amplitude est la méthode la plus pertinente (parmi celles que nous avons testées) pour modéliser le rythme de la parole. En effet, le tempogramme ne nous a pas permis d'obtenir des informations sur des régularités à différents niveaux de la parole. Ce dernier nous a seulement permis d'extraire les régularités syllabiques. À l'inverse, le spectre de modulations de fréquence modélise efficacement les régularités liées aux unités rythmiques larges où se situent les contours intonatifs. En revanche, cette méthode ne permet pas d'étudier les régularités syllabiques. La méthodologie la plus polyvalente selon nous est donc le spectre de modulation d'amplitude. Celui-ci permet en effet de modéliser les différents niveaux prosodiques produits dans la parole. Les résultats sont alors cohérents avec nos observation sur notre corpus de slam. Nous utiliserons donc dans la suite de cette thèse les modulations d'amplitude pour modéliser le rythme. Nous proposerons sans doute à l'avenir de combiner les deux modulations d'amplitude et de fréquence afin d'obtenir une modélisation plus polyvalente.

5

Analyses du rythme de la parole pathologique

Sommaire

Dans le chapitre précédent, nous avons décrit comment nous avons implémenté différentes modélisations du rythme de la parole. Pour cela, nous les avons appliquées à un corpus de slam afin de vérifier que ces modèles permettent bien de mettre en avant les régularités prosodiques de la parole. Suite à ces expérimentations, nous avons pu valider ou rejeter certaines méthodes. Le tempogramme (4.1) ne nous a notamment pas permis de rendre compte efficacement des régularités rythmiques à des niveaux autres que celui de la syllabe. Les spectres de modulations d'amplitude et de fréquence quant à eux ont pu mettre en évidence les divers niveaux prosodiques des énoncés. En revanche, les modulations de fréquences telles que nous les avons implémentées, ne permettaient pas de représenter efficacement les régularités rythmiques supérieures à 4 Hz. Ces fréquences incluent le niveau syllabique qui est un niveau indispensable pour l'étude du rythme en français. Nous nous concentrerons donc davantage sur les modulations d'amplitude, car cette modélisation propose une analyse qui ne repose pas que sur les zones voisées et ne dépend pas d'une extraction de la fréquence fondamentale. Elle permet de rendre compte également des régularités de durées et des

groupements à la base du rythme prosodique. Dans ce chapitre, nous allons appliquer ce modèle à de la parole et plus particulièrement sur nos deux corpus de parole pathologique. Pour rappel, l'un de ces corpus est composé d'enregistrements de personnes atteintes de cancer VADS de la cavité buccale et de l'oropharynx tandis que le second concerne des personnes atteintes de la maladie de Parkinson (MDP). Nous allons donc, dans un premier temps, explorer les particularités rythmiques de ces pathologies en appliquant notre modèle à la tâche de lecture de texte de la chèvre de Monsieur Seguin. Nous allons comparer les résultats obtenus avec les annotations prosodiques que nous avons pu obtenir (cf. section 3.2.2) afin de caractériser au mieux ces pathologies. Par la suite, nous exposons différentes caractéristiques automatiques globales extraites du spectre de modulation d'amplitude. Ces caractéristiques sont ensuite utilisées afin de modéliser la sévérité et l'intelligibilité de la parole afin de voir à quel point le rythme de la parole permet de caractériser ces indices. Nous nous intéresserons enfin à l'extraction automatique de caractéristiques plus précises qui pourraient permettre de décrire le rythme de la parole d'une personne. Ces caractéristiques automatiques sont mises en relation avec des scores prosodiques perceptifs à notre disposition (cf. 3.1.5). À partir des paramètres les plus pertinents, nous allons essayer de caractériser automatiquement les locuteurs en fonction de la structure rythmique de leur parole.

5.1 Le spectre de modulation d'amplitude appliqué à la parole pathologique

Maintenant que nous avons pu voir le comportement du spectre de modulation d'amplitude (EMS) sur nos exemples de slam et que nous sommes capables d'interpréter ses sorties, nous allons pouvoir l'appliquer aux lectures de textes. Cela nous permettra de comparer de manière qualitative comment se comportent les patients par rapport aux sujets témoins. Nous pourrons alors mettre en relation l'EMS avec les annotations prosodiques faites sur le corpus cancer (3.2.2).

5.1.1 Analyse qualitative du rythme de la parole pathologique

Dans un premier temps, il a été nécessaire d'explorer nos corpus de données manuellement afin d'obtenir une meilleure intuition des troubles prosodiques que nous pouvons rencontrer dans le cadre des sujets cancer VADS et Parkinson. Pour cela, nous avons sélectionné arbitrairement quelques locuteurs en nous basant sur leur score de sévérité perceptif (3.1.5). Ainsi, nous avons choisi des personnes à différents degrés de sévérité dans le but de voir les dégradations à différentes étapes des pathologies. Nous avons donc calculé l'EMS dans le cadre de la lecture de texte sur les trois premières phrases du texte dont nous avions proposé une structure prosodique en figure 3.5. La figure 5.1 montre les EMS de trois locuteurs (un sain, un cancer VADS, un Parkinson) sur les deux premières phrases du texte de la chèvre : *"Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres, il les perdait toutes de la même façon"*.

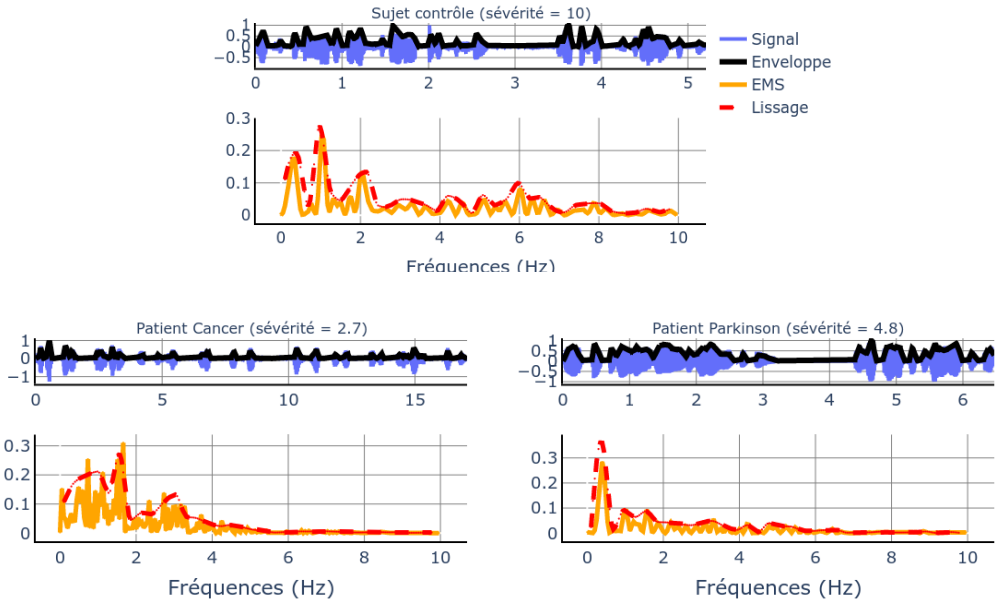

FIGURE 5.1 – Comparaison des EMS de trois individus sur la lecture de l'extrait *"Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres, il les perdait toutes de la même façon."*. Le sujet contrôle (locuteur n^o 033) est en haut, l'EMS du sujet cancer VADS (locuteur n^o 301 ; sévérité = 2.7) est en bas à gauche, celui du patient Parkinson (locuteur n^o 970 ; sévérité = 4.8) est en bas à droite. Les patients sont des personnes avec une sévérité de maladie élevée.

La combinaison de ces deux phrases est particulièrement intéressante car elle propose une bonne diversité des niveaux prosodiques potentiels avec plusieurs groupes de niveau intermédiaire (ip) et deux groupes intonatifs de longueurs différentes. Il sera donc intéressant de regarder si les locuteurs produisent ces différents niveaux avec des durées équivalentes ou non. Sur cette figure, nous pouvons remarquer plusieurs éléments. Tout d'abord, le sujet témoin présente des pics dans son EMS à des fréquences très variables allant de 0,34 Hz (3 sec) à 6 Hz (168 ms) ce qui montre une régularité à plusieurs niveaux tant sur les syllabes que sur les syntagmes intonatifs (IP). Nous retrouvons également des pics à 1 et 2 Hz (1 et 0,5 secondes). Le débit articulatoire de cette personne (nombre de syllabes divisé par la durée de phonation) est environ de 6,1 syllabes par secondes, cette information se retrouve donc dans le dernier pic à 6 Hz. Le pic à 2 Hz indiquerait donc peut être le niveau du syntagme accentuel (trois syllabes environ) et celui à 1 Hz pourrait correspondre au syntagme intermédiaire. Concernant le patient atteint de cancer VADS, nous avons choisi une personne avec une très forte atteinte de sa pathologie. La sévérité de ce patient a été évaluée à 2,7 / 10 par les experts (0 étant la plus forte sévérité possible). Concernant sa lecture, nous pouvons déjà remarquer que son débit de parole est extrêmement lent. Cela peut s'expliquer par de grands troubles de l'articulation. Ces troubles ralentiraient ainsi sa production de parole et forceraient alors le patient à reprendre plus souvent sa respiration. Comme

nous pouvons le voir, ce patient est donc dans l'obligation de segmenter son flux de parole en mots prosodiques (voire parfois en mots). Cependant, malgré cette contrainte physique forte, nous pouvons remarquer que son EMS présente de fortes régularités à plusieurs niveaux. Ainsi, nous retrouvons un pic autour de 0,5 Hz (2 secondes) qui correspondrait à un groupement de plusieurs mots, ainsi qu'un pic autour de 1,6 Hz (625 ms) qui correspond à la récurrence des mots prosodiques. Un dernier pic se situe à 3 Hz et correspond à son débit articulatoire de 2.8 syllabes par secondes (357 ms). Au delà, il n'y a quasiment plus aucune énergie dans le spectre. Cela se traduit à l'écoute par une très mauvaise articulation des syllabes et des phonèmes qui rend l'intelligibilité du patient très faible (score intelligibilité = 1,8). Concernant le patient atteint de la MDP, il présente également quelques troubles articulatoires (plus faibles que pour le patient cancer VADS). En revanche, son débit est davantage similaire au sujet témoin (5 syllabes par secondes). Au niveau de l'EMS, nous pouvons voir de façon assez étonnante que seul le premier pic à 0.34 Hz (2.9 sec) est mis en avant. Cela pourrait donc indiquer une forte régularité du niveau de l'IP mais une régularité beaucoup moins marquée aux autres niveaux aux niveaux prosodiques inférieurs.

Nous n'allons pas ici montrer l'ensemble des fichiers que nous avons pu étudier. Cependant, quelques uns d'entre eux sont visibles dans l'annexe B. Nous pouvons cependant exposer les caractéristiques principales que nous avons pu observer. Concernant la population Parkinson, les atteintes au niveau prosodique nous semblent très variables. Ainsi, de très nombreux patients ne semblent pas de notre point de vue présenter de troubles systématiques de la parole au niveau perceptif. Cette observation est confirmée par les notations expertes de l'intelligibilité qui montrent que 171 patients sur 205 (83%) ont un score d'intelligibilité supérieur à neuf (sur un maximum de dix). Parmi les personnes avec des scores plus faibles, nous avons cependant pu observer quelques particularités. Nous avons pu voir des locuteurs avec un EMS dont l'énergie se situe davantage dans la zone liée aux syllabes. Perceptuellement, ces personnes ont un débit de parole élevé et font de longues pauses entre deux syntagmes. En revanche, nous avons également vu des locuteurs comme sur la figure 5.1 dont l'énergie est essentiellement située vers les basses fréquences. Ce groupe présente davantage un débit plus faible avec des longues pauses ressemblant davantage à ce que nous pouvons trouver dans la population cancer VADS.

Pour les personnes atteintes de cancer VADS, la diversité rythmique est plus grande. Tout d'abord, les personnes avec une intelligibilité élevée sont moins nombreuses avec 40 personnes sur 87 avec un score supérieur à neuf (46%). Parmi les personnes avec une intelligibilité (et sévérité) plus faible, nous avons pu observer un phénomène particulier. Les EMS des patients suivent une tendance telle que l'énergie située dans la zone supérieure à 4 Hz semble corrélée négativement avec celle dans la zone inférieure à 4 Hz. En effet, plus l'énergie dans les hautes fréquences est élevée, plus celle des fréquences plus faibles est basse et inversement. Nous supposons que ce phénomène est dû aux troubles articulatoires des patients. Notre hypothèse est la suivante : Plus les patients sont atteints de trouble sévère de l'articulation, plus ils le compensent par une forte régularité rythmique des unités prosodiques supérieures à la syllabe (Vaysse

et collab., 2021). Afin de vérifier cette théorie, nous avons réalisé une annotation des niveaux prosodiques de la lecture sur une sélection de 10 patients et 10 sujets contrôles (voir 3.2.2).

5.1.2 Les pics de l'EMS en lien avec les niveaux prosodiques

Afin de tester nos observations, nous avons donc voulu superposer nos spectres de modulation d'amplitude avec les annotations prosodiques que nous avons présentées dans la partie 3.2.2. Ces annotations indiquent les durées de plusieurs unités rythmiques dont : le niveau de la syllabe, du mot prosodique (pw), du syntagme accentuel (ap), du syntagme intermédiaire (ip) et du syntagme intonatif (IP). Un exemple d'annotation est donné dans la figure 3.10 (p.64). Ces informations ont été produites sur 10 patients cancer VADS et 10 locuteurs sains. Nous n'en avons en revanche à ce jour pas encore produit pour les patients MDP.

Afin de superposer ces informations à nos EMS, nous avons décidé d'étudier les durées moyennes de chaque unité pour les trois premières phrases de la chèvre de Monsieur Seguin (voir figure 3.5 p.70 pour l'annotation prosodique de ces phrases). À partir de ces durées, nous avons estimé des intervalles de fréquences qui correspondent aux différentes unités. Les bornes de ces intervalles ont été constituées de telle sorte :

$$b_{inf} = \frac{1}{m + \sigma\sqrt{2}} \quad b_{sup} = \frac{1}{m - \sigma\sqrt{2}}$$

Avec m la durée moyenne de l'unité et σ son écart type. La formule $m \pm \sigma\sqrt{2}$ est un intervalle qui inclue au minimum 50% des durées de cette unité en considérant que ces durées suivent une distribution quelconque d'après l'inégalité de Chebyshev (Ghosh, 2002). Nous n'avons pas utilisé $m \pm \sigma$ car suite à des tests de Shapiro-Wilk, nous avons remarqué que pour quelques patients les distributions de durées ne suivent pas une loi normale. Cependant, la plupart des tests ne permettant pas de réfuter la normalité des données, nous pouvons supposer que notre intervalle englobe bien plus que 50% des données ($m \pm \sigma\sqrt{2}$ englobe 68% des données pour une loi normale).

Un exemple d'EMS superposé aux informations prosodiques est visible sur la figure 5.2 où nous avons pris l'EMS d'un sujet témoin (le même que sur la figure 5.1) sur les deux premières phrases du texte "*Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon*".

Sur cet exemple, nous pouvons donc voir que les trois premiers pics de régularité de l'EMS correspondent très bien avec les niveaux de l'IP, l'ip et l'ap annotés manuellement. On constate alors une bonne régularité de ces trois niveaux ainsi que des pics de régularité syllabique à diverses fréquences supérieures à quatre hertz. Le niveau du mot prosodique n'apparaît pas clairement dans notre exemple car le court texte lu ne permet pas de distinguer facilement d'un point de vue syntactico-prosodique le niveau de l'ap et du pw, notamment chez les patients relativement fluents. Du point de vue pathologique, nous n'avons pu comparer l'EMS et les annotations prosodiques que pour les personnes atteintes de cancer VADS pour lesquelles nous avons à notre

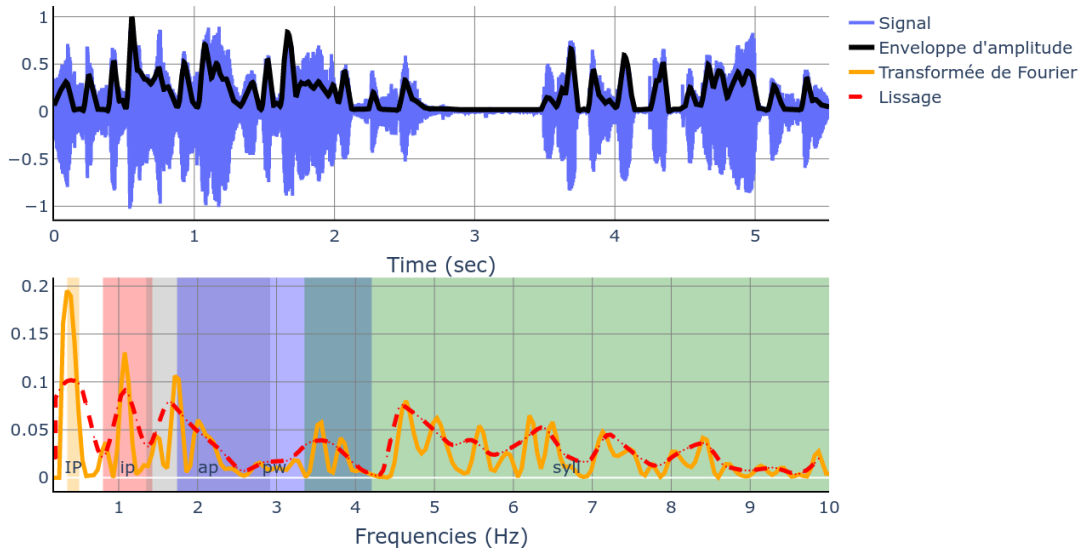

FIGURE 5.2 – EMS d'un locuteur sain (locuteur n^o 033) sur l'extrait "*Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres, il les perdait toutes de la même façon*". Les intervalles des niveaux prosodiques sont indiqués en couleur : orange pour l'IP, rouge pour l'ip, bleu pour le pw, gris pour l'ap et vert pour la syllabe.

disposition des annotations. Prenons par exemple le cas de la même personne cancer VADS que sur la figure 5.1 qui produisait une segmentation mot par mot de l'énoncé avec une articulation très détériorée. L'EMS sur la première phrase est visible sur la figure 5.3. Sur cet exemple, les intervalles de l'ap et du pw sont superposés car la personne segmente son discours en mots prosodiques. L'ap et le pw correspondent donc ici aux mêmes unités. Nous pouvons alors observer de nouveau un large pic dans l'intervalle de l'ip entre 0,6 Hz (1,7 sec) et 1,1 Hz (909 ms). De même, un autre pic marque le niveau de l'ap/pw à 1,6 Hz (625 ms). Le pic lié aux syllabes est à une fréquence bien plus basse que pour la personne saine de par un débit de parole bien plus faible. L'IP en revanche, n'est pas repérée comme étant régulière. Il est en effet plus difficile de produire des régularités dans des unités aussi larges avec un débit de parole aussi peu élevé (la durée moyenne de ses IP est de 8 secondes ici).

Afin de vérifier si la correspondance entre pics et unités prosodiques est bien réelle, nous pouvons également regarder sur la figure 5.4 qui représente l'EMS d'un patient encore plus touché par sa pathologie avec un score de sévérité de 1,3 (1,8 en intelligibilité). Ce locuteur présente également des troubles articulatoires qui impactent fortement son intelligibilité. Tout comme sur le locuteur précédent, nous pouvons constater une forte régularité des unités intermédiaires et accentuelles avec cette fois ci des unités légèrement plus larges. Là où le premier locuteur segmentait sa parole en mots, celui-ci la segmente en groupes de deux ou trois mots. Nous pouvons voir sur le signal une segmentation très régulière avec des groupements de mots d'environ 1,5 secondes. Ces groupements dans notre annotation sont considérés comme étant des

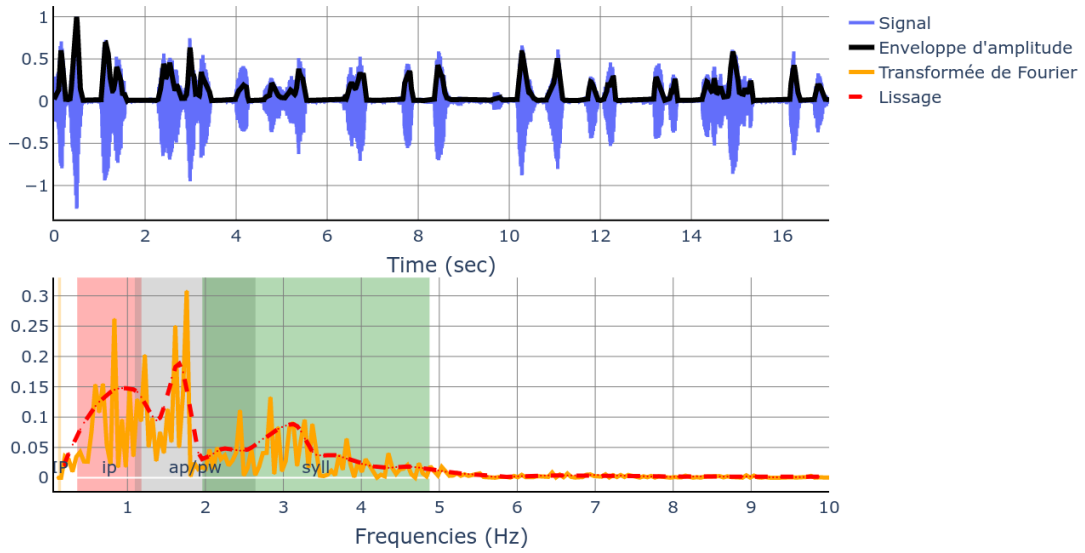

FIGURE 5.3 – EMS d'un locuteur cancer VADS (locuteur n^o 304 ; sévérité = 2,6) sur les phrases "*Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon.*". Les intervalles des niveaux prosodiques sont indiqués en couleur : rouge pour l'ip, gris pour l'ap (ici équivalent au pw) et vert pour la syllabe.

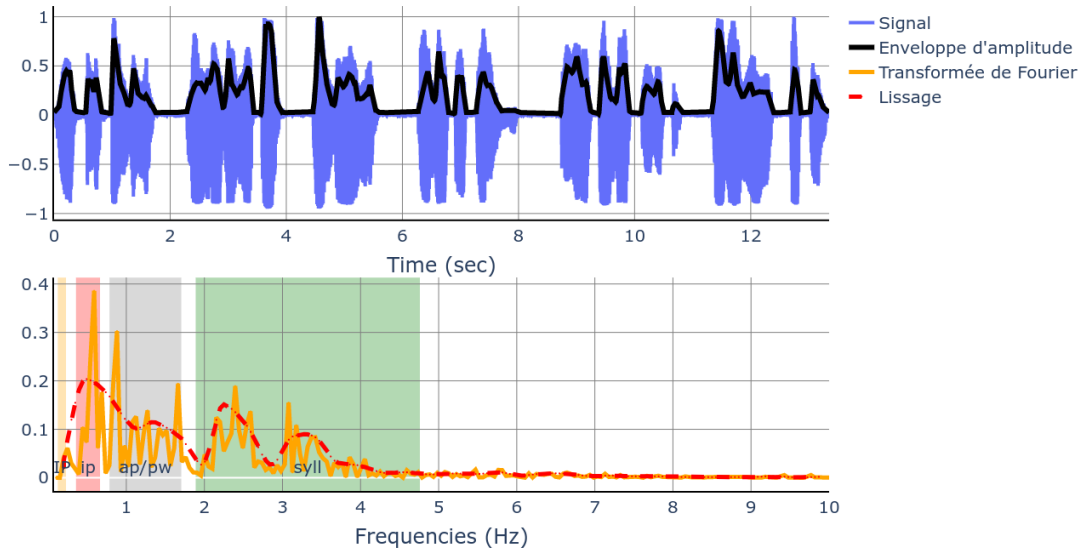

FIGURE 5.4 – EMS d'un locuteur cancer VADS (locuteur n^o 308 ; sévérité = 1,3) sur les phrases "*Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon*". Les intervalles des niveaux prosodiques sont indiqués en couleur : rouge pour l'ip, gris pour l'ap (ici équivalent au pw) et vert pour la syllabe.

unités intermédiaires (ip). Nous pouvons alors observer cette régularité sur l'EMS avec un pic à 0,59 Hz (1,7 sec). En plus de ce pic, nous avons également la régularité en ap (ou en pw) qui ressort à 1,4 Hz (714 ms) étant donné que la plupart de ses ip sont divisées en deux mots. Une régularité syllabique sur deux pics est également observée. Les durées associées à ces pics pourraient alors correspondre respectivement aux syllabes accentuées et inaccentuées de cet énoncé. Ce phénomène ne se retrouve cependant que rarement sur les EMS d'autres locuteurs. De fait, ce locuteur qui présente des difficultés articulatoires importantes semble privilégier la fluence rythmique à cause de sa faible précision articulatoire.

Les trois exemples que nous avons montré ici nous semblent être des exemples prototypiques de stratégies rythmiques que nous retrouvons dans le corpus de cancer VADS. Ils ne représentent cependant pas tous les types de paroles de ce corpus. En effet, bien que pour ces exemples l'alignement entre les pics et les unités prosodiques soit très bon, il n'est cependant pas systématiquement aussi parfait. Il arrive alors parfois de ne pas trouver de pics pour certaines unités. Cette dernière observation nous mène à penser que dans certaines conditions, l'EMS n'est pas capable de rendre compte précisément de toutes les unités prosodiques mises en jeu. Cette modélisation fournit néanmoins une bonne estimation des unités rythmiques et de leurs durées. D'autres exemples d'EMS superposés avec les annotations prosodiques sont disponibles dans l'annexe.

Pour résumer, nous sommes parvenus à déceler différentes stratégies de production de la parole au sein de la population des patients atteints de cancers VADS. Les patients atteints de troubles articulatoires sévères dotés d'une mauvaise fluence ont un débit de parole faible de sorte à améliorer au maximum leurs chances de produire des syllabes compréhensibles. Cette baisse de débit s'accompagne alors d'une structuration rythmique particulière où le niveau prosodique dominant est le pw. La régularité de ce niveau est alors très forte avec des durées d'un pw à un autre très similaires. Sur l'EMS, cela se traduit par un petit nombre de hauts pics situés dans la première moitié du spectre et une seconde moitié avec une énergie quasi nulle. Il existe également des patients avec des troubles articulatoires sans perte de fluence. Ces personnes présentent alors des regroupements prosodiques très réguliers à des niveaux supérieurs comme l'ap ou l'ip. Par ailleurs, leur EMS ne contient que peu d'énergie dans la seconde moitié du spectre ([5-10] Hz) et un ou deux pics dominants au niveau de l'ap ou l'ip. Enfin, nous avons des patients et sujets sains qui n'ont pas de troubles particuliers de l'articulation mais qui en contrepartie ne produisent pas toujours leurs unités prosodiques de façon aussi régulière que les autres patients. Ceci résulte pour ces derniers en un EMS où de nombreux pics apparaissent sur l'ensemble de la bande de fréquence [0-10] Hz.

5.2 Vers une prédiction de l'intelligibilité de la parole pathologique

Maintenant que nous avons vu comment se comporte l'EMS sur des enregistrements de patients et de sujets contrôles, nous allons pouvoir déterminer et extraire des indices automatiques qui permettront de caractériser les spectres de modulations des locuteurs. Dans un premier temps, nous allons nous focaliser sur l'extraction de caractéristiques générales de l'EMS à l'aide de paramètres ne se focalisant pas sur des niveaux prosodiques précis. Ces paramètres concernent des mesures d'énergie dans de larges bandes de fréquences regroupant plusieurs niveaux prosodiques ou encore des statistiques générales comme l'amplitude ou la fréquence des principaux pics de l'EMS. Nous souhaitons ainsi savoir si ces paramètres permettent ou non de modéliser en partie l'intelligibilité et/ou la sévérité des patients. En fonction des résultats, nous pourrons alors établir à quel degré le rythme de la parole influence l'intelligibilité d'une personne.

5.2.1 La prédiction de l'intelligibilité par des caractéristiques rythmiques

Dans cette partie, nous allons donc nous focaliser sur l'extraction automatique de descripteurs de l'EMS. Nous essayerons par la suite de prédire l'intelligibilité et la sévérité des patients via des algorithmes d'apprentissage automatique. Pour cela, nous nous sommes principalement inspirés de l'étude réalisée par Liss et collab. (2010) dans laquelle ils extraient des statistiques sur l'EMS dans le but de prédire différents types de dysarthries. Liss et collab. (2010) se basent également sur une tâche de lecture de texte. Ils extraient l'EMS sur l'ensemble du fichier audio et calculent huit paramètres sur cet EMS :

— La fréquence et l'amplitude des deux plus hauts pics de l'EMS (quatre paramètres)
— L'énergie dans la bande de fréquence entre 3 et 6 Hz (166 à 333 ms) qui englobe une grande partie des durées des syllabes en anglais Arai et Greenberg (1997); Liss et collab. (2010).
— L'énergie en dessous de 4 Hz
— L'énergie entre de 4 et 10 Hz. Les bandes [0-4] et [4-10] Hz ont été choisies car ce sont celles qui empiriquement sont les moins corrélées entre elles Liss et collab. (2010). Les informations contenues dans ces deux bandes sont donc complémentaires.
— Le ratio d'énergie entre les bandes 0-4 et 4-10 Hz

Dans notre cas, nous avons utilisé les fichiers de lecture de la chèvre de Monsieur Seguin (cf. section 3.1.4). En revanche, à l'inverse de l'étude de Liss et collab. (2010), nous n'avons pas réalisé l'extraction de l'EMS sur l'ensemble du signal d'un coup, mais nous avons effectué ce calcul sur une fenêtre glissante. Nous avons ensuite extrait les

paramètres de l'EMS sur chacune de ces fenêtres. Ainsi, notre procédure peut être décrite comme ceci :

1. Extraction d'une fenêtre temporelle de cinq secondes du signal
2. Calcul de l'EMS sur cette fenêtre
3. Extraction de paramètres (amplitude, fréquence, énergies)
4. Sauvegarde des paramètres
5. Décalage temporel d'une demi seconde et réitération des étapes 1 à 5 jusqu'au bout du fichier audio
6. Calcul de statistiques (moyenne, écart-type, asymétrie, aplatissement) de chaque paramètre

Cette méthodologie présente selon nous plusieurs avantages. Tout d'abord, l'extraction sur des fenêtres de cinq secondes nous permet d'analyser le rythme sur des durées moins importantes et de pouvoir détecter des changements de régularité d'une phrase à une autre. De plus, l'extraction sur des fenêtres glissantes nous permet d'analyser par exemple l'évolution de l'énergie entre 4 et 10 Hz au cours du temps. Cette évolution peut alors être caractérisée par sa valeur moyenne, mais également par des informations sur ses variations comme son écart type ou ses moments d'ordres supérieurs comme l'asymétrie (skewness) ou son aplatissement (kurtosis). Dans les paramètres extraits, nous avons également utilisé l'énergie [0,5-4] Hz à la place de [0-4] Hz étant donné que sur une fenêtre de temps de 5 secondes, les fréquences très basses ne sont pas pertinentes. Un exemple de comparaison entre l'évolution du ratio d'énergie [0,5-4] / [4-10] Hz chez une personne atteinte d'un cancer VADS et d'une personne saine est disponible en figure 5.5. Sur cette figure, nous pouvons voir que

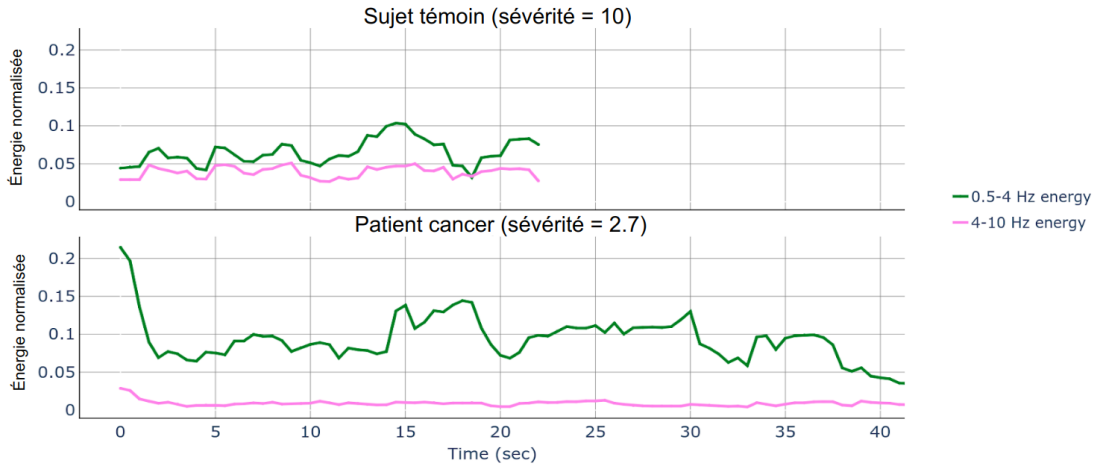

FIGURE 5.5 – Courbes de l'énergie dans les bandes de fréquences [0,5-4] Hz (en vert) et [4-10] Hz (en rose) en fonction du temps. Deux locuteurs sont représentés, un sain en haut (locuteur n^o 032) et un atteint de cancer VADS en bas (locuteur n^o 301).

l'énergie dans la bande [4-10] Hz du patient est plus faible que celle du sujet contrôle. Cela pourrait alors témoigner de la faible articulation des phonèmes et des syllabes

ou bien plus simplement par des différences de débit de parole (nous rediscuterons de cela dans la partie 5.2.2). En revanche, l'énergie [0,5-4] Hz est légèrement plus forte chez le patient. À partir de ce genre de courbes, nous extrayons les quatre statistiques évoquées précédemment (moyenne, écart-type, asymétrie, aplatissement) afin de quantifier les informations rythmiques du signal.

Nous avons donc une méthode capable d'extraire plusieurs caractéristiques rythmiques automatiquement pour chaque locuteur. À partir de ces paramètres et des annotations cliniques expertes, il est donc possible d'utiliser des algorithmes d'apprentissage automatique afin de réaliser une estimation de l'intelligibilité et de la sévérité à partir du rythme de la parole. Pour cela, nous avons au total 32 paramètres par locuteur (les huit paramètres listés au début de cette partie multipliés par quatre statistiques). Nos calculs ont été effectués séparément sur le corpus de parole cancer VADS et sur celui de parole Parkinson afin de voir si les résultats étaient différents entre ces deux pathologies. Afin de réaliser la prédiction des scores d'intelligibilité et de sévérité, nous avons utilisé une méthode d'apprentissage dit supervisé. Le principe de ce type de méthode est de fournir un jeu de *données d'entraînement* (X, y) avec X les paramètres (dans notre cas les paramètres extraits de l'EMS) et y la "vérité terrain" qui ici représente un score expert. Le modèle d'apprentissage supervisé va alors apprendre la correspondance entre les paramètres X et les valeurs attendues y. À partir de cette correspondance, nous lui fournissons un nouveau jeu de *données de test* (X', y') qu'il n'a jamais vu afin de vérifier s'il est capable de fournir une estimation correcte de y'.

Dans notre expérience, le corpus de parole cancer VADS étant relativement petit (113 enregistrements), nous avons décidé d'utiliser une méthodologie d'évaluation particulière. Afin d'évaluer les performances de l'algorithme, nous avons réalisé une validation croisée avec une méthode "tous sauf un" (*Leave One Out*). Dans cette méthodologie, le jeu d'entraînement (X, y) est composé de tous les locuteurs à l'exception d'un d'entre eux et le jeu de test (X', y') est composé du locuteur restant. Le modèle n'ayant jamais vu le dernier locuteur, nous pouvons comparer la prédiction $\hat{y}'$ du modèle avec la véritable valeur y'. En répétant cette opération jusqu'à avoir testé le modèle sur tous les locuteurs, nous obtenons une évaluation robuste de la qualité du modèle. Une illustration de cette méthodologie est disponible en figure 5.6.

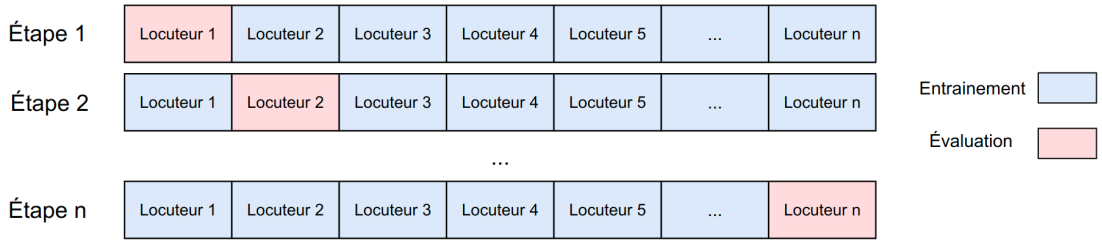

FIGURE 5.6 – Illustration de la méthode de validation croisée "tous sauf un" (*Leave One Out*). Chaque rectangle représente les paramètres d'un locuteur. À chaque étape, on entraîne le modèle d'apprentissage sur les locuteurs bleus et on l'évalue sur le locuteur rouge.

Afin de réaliser l'apprentissage, nous avons utilisé un régresseur à vecteurs de support (*Support Vector Regressor*, SVR) (Drucker et collab., 1997). Le SVR est une adaptation de l'algorithme SVM (Machine à Vecteur Support) pour la régression (prédiction d'une variable continue). Cet algorithme consiste à trouver une combinaison linéaire des paramètres (e.g moyenne de l'énergie [4-10] Hz) dont les valeurs diffèrent au maximum d'un écart ϵ par rapport aux valeurs de référence (e.g sévérité), pour toutes les données d'entraînement. Lorsque cela n'est pas réalisable, des variables augmentant le degré de liberté sont introduites pour résoudre le problème. Cette modélisation est particulièrement performante sur des jeux de données de petite taille. Cependant, notre nombre de paramètres (32) est relativement haut par rapport au nombre d'enregistrements pour le corpus cancer VADS (113). C'est la raison pour laquelle nous avons utilisé en amont de l'apprentissage du SVR, une méthode de réduction de paramètres. Cela permet de réduire le nombre de paramètres d'un modèle de sorte à ne conserver que les plus pertinents. Cette réduction de paramètre permet alors d'éviter le phénomène de sur-apprentissage. C'est un phénomène dans lequel le modèle généraliserait mal son apprentissage réalisé sur le jeu d'entraînement et qui fournirait donc de mauvais résultats sur le jeu de test. La méthode de sélection de paramètres que nous avons choisie est une simple régression linéaire à laquelle on a ajouté une régularisation LASSO (*Least Absolute Shrinkage and Selection Operator*) (Tibshirani, 1996). Le principe de base d'une régression linéaire est de réaliser une combinaison linéaire des paramètres (X) de sorte à générer une prédiction la plus proche possible de y. La régularisation LASSO ajoute à cet objectif une contrainte dans laquelle les coefficients associés à chaque paramètre doivent être les plus petits possibles. Ainsi, les paramètres (e.g fréquence des pics, amplitude...) les moins pertinents sont associés à des coefficients nuls. Donc en éliminant ces paramètres, nous pouvons réduire fortement leur nombre et les fournir au SVR.

Une fois la réduction de paramètres réalisée, nous obtenons alors un nouvel ensemble de paramètres pertinents pour le score à prédire. Par exemple, dans le cadre de la prédiction de l'intelligibilité sur le corpus cancer VADS, la sélection a retenu au total cinq paramètres sur les 32 initiaux :

— La moyenne du ratio entre l'énergie [0,5-4] et [4-10] Hz
— L'asymétrie de la fréquence du plus haut pic
— La fréquence moyenne du second plus haut pic
— L'énergie moyenne entre 0,5 et 4 Hz
— L'aplatissement de l'énergie entre 3 et 6 Hz

En ne fournissant que ces cinq paramètres au SVR, nous obtenons alors de bons résultats dans les tâches de prédiction de l'intelligibilité et de la sévérité. Sur le corpus cancer VADS, les valeurs d'intelligibilité prédites sont corrélées à 0,73 avec l'intelligibilité cliniques de référence. Pour la sévérité, cette corrélation monte à 0,8. Des nuages de points comparant nos prédictions aux valeurs cliniques sont visibles sur la figure 5.7. Nous pouvons observer que notre modèle prédit mieux la sévérité que l'intelligibilité. Ceci était attendu étant donné que la mesure de sévérité prend davantage en compte les variabilités prosodiques. Ainsi, pour les personnes avec des scores faibles

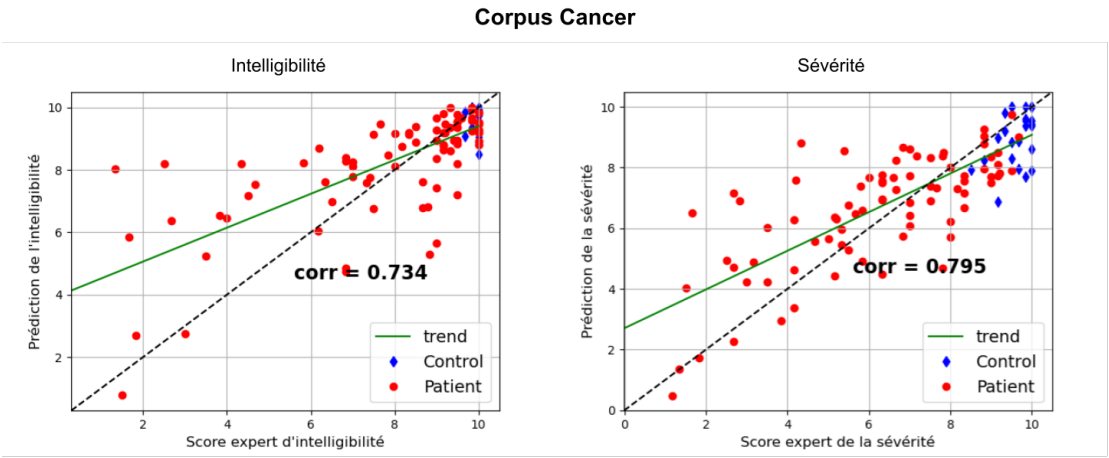

FIGURE 5.7 – Prédiction automatique de l'intelligibilité (à gauche) et de la sévérité (à droite) en fonction des évaluations expertes respectives sur le corpus de parole cancer VADS. Les patients sont en bleu (losanges) et les sujets témoins en rouge (ronds).

d'intelligibilité, notre modèle prédit des scores souvent bien plus hauts. Au niveau de la sévérité en revanche, nous trouvons également cette sur-évaluation de certains patients, mais de façon moins systématique. Afin d'évaluer la pertinence de nos prédictions, nous avons calculé également deux métriques en plus de la corrélation : l'erreur moyenne absolue (MAE) et l'erreur moyenne quadratique (RMSE). Ces deux métriques se calculent comme suit :

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2} \quad MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|$$

Où N est le nombre de locuteurs, y_i le score (sévérité / intelligibilité) perceptif clinique du i-ème locuteur et $\hat{y}_i$ la prédiction de ce score par notre SVR. Le modèle d'intelligibilité a une MAE de 1,01 pour l'intelligibilité contre 1,16 pour la sévérité. Ce qui signifie qu'en moyenne notre prédiction de la sévérité diffère de 1,16 par rapport au score de référence. L'erreur moyenne est donc en faveur du score d'intelligibilité, cependant pour la RMSE, cette tendance s'inverse avec 2,36 pour l'intelligibilité et 2,25 pour la sévérité. Cette différence entre MAE et RMSE signifie simplement que nos prédictions d'intelligibilité sont en moyenne plus proches (MAE plus faible) de la vérité terrain, mais que certaines erreurs sont relativement importantes (RMSE plus haute). Cela se retrouve comme expliqué précédemment dans la prédiction des scores des personnes les plus atteintes où le modèle de sévérité est plus pertinent que celui de l'intelligibilité. Cette étude sur la prédiction de l'intelligibilité pour le corpus cancer VADS a fait l'objet d'une publication dans la conférence internationale Interspeech (Vaysse et collab., 2021). Dans cet article, nous n'avons cependant pas exposé nos résultats sur le corpus Parkinson.

Concernant le corpus Parkinson, nous avons réalisé la même expérience. La corrélation entre nos prédictions d'intelligibilité et la vérité terrain sont de 0,68 et pour la sévérité, elle est de 0,63. Les prédictions sont visibles sur la figure 5.8. Étonnamment,

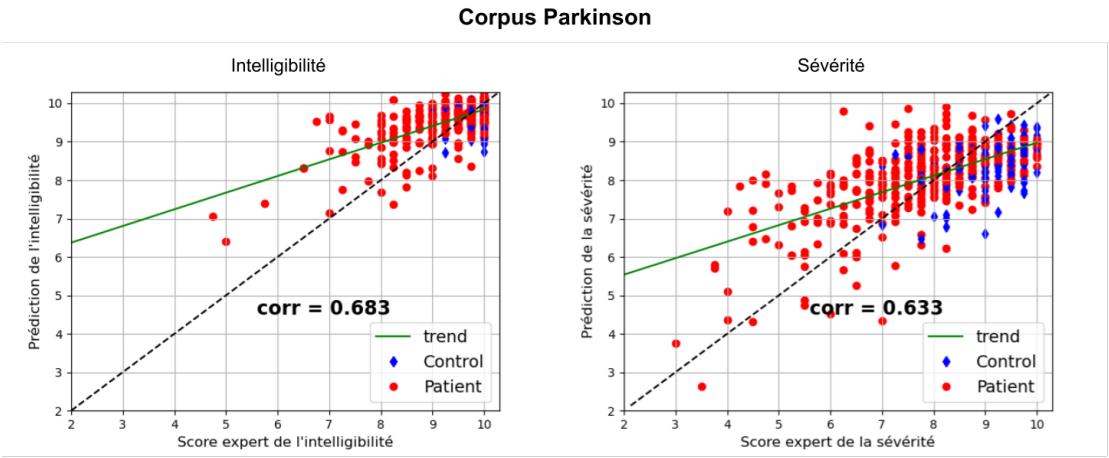

FIGURE 5.8 – Prédiction automatique de l'intelligibilité (à gauche) et de la sévérité (à droite) en fonction des évaluations expertes respectives sur le corpus de parole Parkinson. Les patients sont en bleu (losanges) et les sujets témoins en rouge (ronds).

les résultats obtenus semblent inversés par rapport au corpus cancer VADS, avec une corrélation pour l'intelligibilité de 0,68 supérieure à celle pour la sévérité (0,63). Au niveau de la MAE, elle est de 0,27 pour l'intelligibilité et 0,77 pour la sévérité. Cependant, ces très bons résultats sont en partie dus au fait que ce corpus contient de nombreuses personnes avec des scores cliniques élevés. Il est en effet plus simple pour notre modèle de prédire des valeurs élevées étant donné que le corpus n'est pas équitablement réparti au niveau des scores perceptifs. Ainsi, si nous regardons la MAE seulement sur les personnes ayant un score strictement inférieur à 8, nous obtenons une MAE de 1,57 sur l'intelligibilité et 1,0 pour la sévérité. Notre modèle est donc finalement bien meilleur pour prédire la sévérité des patients avec une sévérité de la maladie modérée ou forte.

Au delà de la qualité de nos prédictions, nous nous sommes intéressés aux paramètres qui ont été les plus utilisés par notre SVR. Quel que soit le corpus ou le score à prédire (intelligibilité / sévérité), le paramètre le plus utilisé par nos modèles est le ratio entre les énergies des bandes de fréquences [0,5-4] et [4-10] Hz. Ainsi, nous avons pu constater que plus ce ratio est élevé, plus les scores cliniques sont proches de 0. Cela pourrait donc renforcer notre hypothèse faite dans la section 5.1.1 selon laquelle, plus une personne est touchée par sa maladie, plus ses qualités articulatoires baissent (énergie [4-10] Hz faible) et plus sa régularité d'unités prosodiques supérieures à la syllabes (pw, ap, ip ou IP) est forte (énergie [0-4] Hz élevée). Bien que nos résultats semblent en adéquation avec notre hypothèse, il pourrait également y avoir une autre explication à l'importance de ce ratio. Il se pourrait plus simplement que les personnes avec des symptômes sévères puissent avoir un débit de parole plus faible que les autres. Ainsi, un débit plus faible décalerait l'énergie de l'EMS vers la gauche (vers les fréquences plus faibles). Ce décalage provoquerait alors une augmentation de notre ratio d'énergie. Par exemple, si une personne possède un débit syllabique de 4,5 syllabes par secondes, il est probable de trouver de l'énergie dans la zone 4,5 Hz

sur l'EMS de cette personne. En revanche, si une autre personne a un débit de 3,5 Hz, cette énergie correspondant à la régularité syllabique se retrouve plus à gauche dans l'EMS. Utiliser une frontière fixe de 4 Hz pour notre calcul de ratio n'est donc potentiellement pas la meilleure stratégie.

5.2.2 Neutralisation du débit

Les pathologies que nous étudions peuvent avoir un impact sur le débit de parole. Le débit peut être étudié via deux paramètres : le débit de parole et le débit articulatoire. Le débit de parole se calcule comme le nombre de syllabes prononcées, divisé par la durée totale de parole. Tandis que le débit articulatoire se calcule comme le nombre de syllabe, divisé par la durée de parole en excluant les pauses. Le débit articulatoire nous fournit donc une information des durées syllabiques sans pauses. Pour cela, de nombreuses méthodes existent et sont performantes sur de la parole typique (Narayanan et Wang, 2005; Wang et Narayanan, 2007; De Jong et Wempe, 2009). En revanche le contexte atypique de la parole pathologique rend cette tâche plus difficile. Nous avons donc voulu tester deux méthodes simples afin de voir comment extraire simplement et efficacement les informations de débit de parole et articulatoire. La première est une méthode que l'on retrouve chez Pellegrino et collab. (2004) et De Jong et Wempe (2009) : les débits sont calculés en recherchant des noyaux vocaliques au sein du signal et ainsi en déduire le nombre de syllabes prononcées. Cette méthode très efficace sur la parole typique risque cependant de rater des noyaux vocaliques chez des patients avec de forts troubles de l'articulation. La seconde méthode, plus simple, consiste à compter le nombre de syllabes théorique dans notre texte (74) et d'utiliser un outil de détection d'activité vocale comme WebRTC-VAD (Google, 2011) pour mesurer la durée des zones de parole. Afin de vérifier la pertinence de ces algorithmes pour estimer le débit de parole, nous avons utilisé les dix fichiers du corpus cancer VADS qui ont été annotés au niveau de leur structure hiérarchique (voir 3.2.2). Ces fichiers ont été annotés et corrigés manuellement au niveau de la syllabe et fournissent donc un point de comparaison fiable. Bien que le nombre de fichiers soit faible, il comporte des patients VADS dont les traitements (localisation de l'ablation, chimiothérapie, radiothérapie...) sont divers et donc induisent des scores de sévérité variés. Nous pensons donc que les résultats seront généralisables.

Nous avons donc estimé le débit de parole et articulatoire de ces dix fichiers avec nos deux méthodologies et nous avons trouvé qu'en moyenne l'erreur absolue d'estimation (MAE) du débit de parole est de 0,31 pour la première méthode. Pour la méthode basée sur le nombre théorique de syllabes, la MAE est de seulement 0,09. Pour le débit articulatoire, on obtient une MAE de 0,71 pour la première et 0,24 pour la seconde. Nous avons donc décidé de continuer à utiliser cette méthode de calcul naïve bien que nous sommes conscient que si le locuteur se trompe dans sa lecture ou répète des mots, les valeurs ne seront pas parfaites.

Les patients atteints de la maladie de Parkinson ne semblent pas montrer de troubles majeurs du débit de parole avec un débit de parole moyen de 2,83 syll/sec et un

débit articulatoire de 4,34 syll/sec contre 2,91 syll/sec et 4,28 syll/sec chez les sujets contrôles. Sur le corpus cancer VADS, cette différence est davantage marquée avec un débit de parole de 2,25 syll/sec et un débit articulatoire de 3,48 chez les patients contre 2,87 et 4,37 chez les contrôles. Nous avons également calculé les corrélations entre le débit de parole et la sévérité, l'intelligibilité et notre ratio d'énergie. Ces corrélations sont visibles sur la figure 5.9.

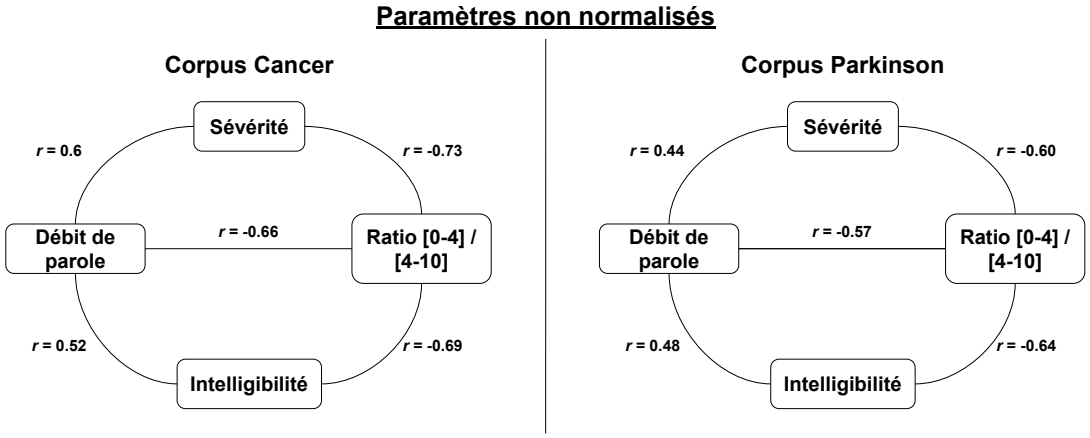

FIGURE 5.9 – Schéma des corrélations de Pearson entre les différents paramètres automatiques et les scores perceptifs en fonction du corpus de parole pathologique.

Nous pouvons ainsi voir que le débit de parole est un élément non négligeable de tous ces paramètres et qu'il est particulièrement corrélé à notre ratio d'énergies, notamment pour les VADS. Cependant, ce qui nous intéresse ici est d'étudier l'impact du rythme de la parole sur l'intelligibilité et la sévérité indépendemment du débit de parole. Afin d'atténuer cette corrélation, nous avons donc voulu essayer une méthode de calcul différente en extrayant cette fois ci les énergies de l'EMS non pas sur des bandes de fréquences fixes, mais sur des bandes adaptées à chaque locuteur. Pour cela, nous avons remplacé la valeur pivot de 4 Hz de nos calculs d'énergies pour la remplacer par la valeur du débit articulatoire. Ce débit étant généralement proche de 4 Hz, l'information contenue dans ces nouveaux paramètres devrait donc être proche de celle initiale. Ainsi, en utilisant les bandes de fréquences [0-ar] et [ar-10] Hz (avec ar le débit articulatoire) à la place de [0-4] et [4-10] Hz, nous espérons conserver les informations d'énergies liées aux syllabes dans la zone [ar-10]. Ceci nous permettrait d'obtenir des paramètres normalisés cohérents avec leurs version non-normalisées qui contribuaient fortement à la prédiction des scores cliniques. Pour vérifier cela, nous avons comparé les corrélations entre les bandes d'énergies fixes ([0-4] / [4-10] Hz), les bandes adaptées aux locuteurs et le débit de parole. La corrélation entre le débit et le ratio d'énergie passe alors sur le corpus cancer VADS de -0,66 avec l'ancienne méthode à 0,19 avec la nouvelle. L'énergie dans la bande [0-4] Hz est quant à elle corrélée à 0,89 à l'énergie dans la bande [0-ar] Hz. La bande [4-10] est corrélée à 0,87 avec la bande [ar-10]. Nous parvenons alors à conserver la majorité de l'information pertinente tout

en diminuant fortement l'influence du débit de parole. Ces nouvelles corrélations sont résumées dans la figure 5.10.

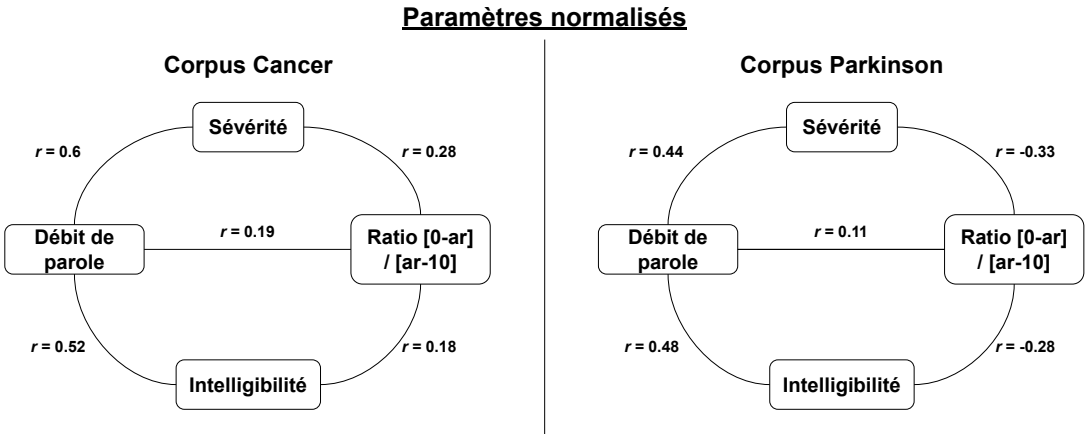

FIGURE 5.10 – Schéma des corrélations de Pearson entre les différents paramètres normalisés par le débit de parole et les scores cliniques en fonction du corpus de parole pathologique.

Nous avons donc pu réitérer la modélisation des scores cliniques avec ces nouveaux paramètres en enlevant de plus les deux paramètres de fréquence des plus hauts pics. Ces paramètres fournissaient également des informations sur le débit de parole mais n'étaient de toute façon pas très utilisés par notre SVR. En réalisant cela, nous obtenons des résultats beaucoup moins bons comme nous pouvons le voir sur les figures 5.11 et 5.12.

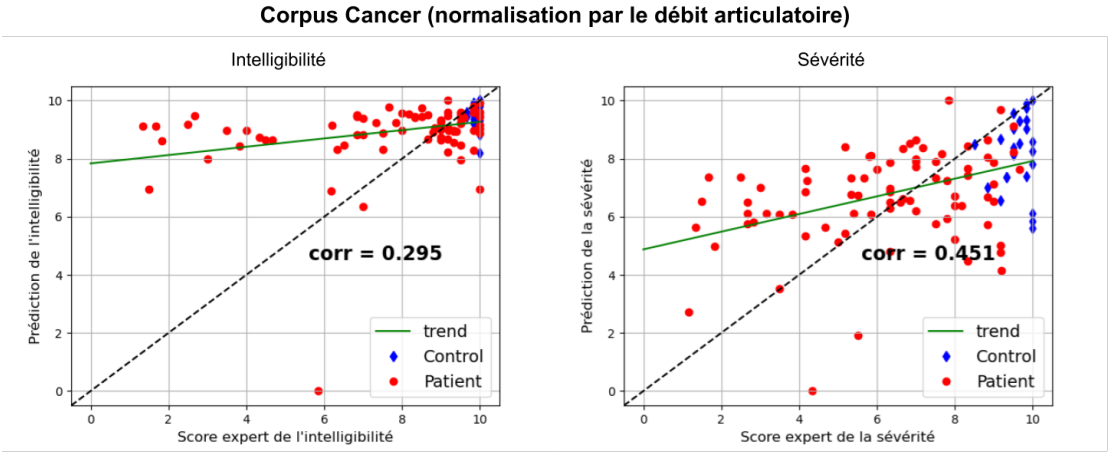

FIGURE 5.11 – Prédiction automatique de l'intelligibilité (à gauche) et de la sévérité (à droite) en fonction des évaluations expertes respectives sur le corpus de parole Parkinson. Les modèles ont été entraînés en atténuant la corrélation des paramètres avec le débit de parole. Les patients sont en bleu (losanges) et les sujets témoins en rouge (ronds).

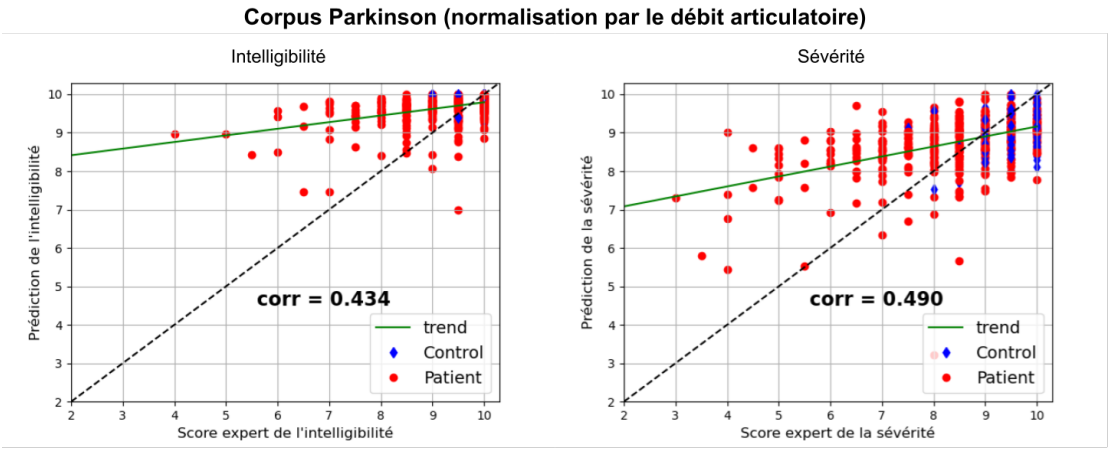

FIGURE 5.12 – Prédiction automatique de l'intelligibilité (à gauche) et de la sévérité (à droite) en fonction des évaluations expertes respectives sur le corpus de parole Parkinson. Les modèles ont été entraînés en atténuant la corrélation des paramètres avec le débit de parole. Les patients sont en bleu (losanges) et les sujets témoins en rouge (ronds).

Nos prédictions sont alors beaucoup moins pertinentes, notamment chez les personnes fortement atteintes par leur maladie. Pour ces personnes, nous pouvons donc supposer que notre modèle se basait principalement sur des informations de débit de parole. La corrélation entre nos prédictions et la vérité terrain n'étant néanmoins pas totalement supprimée, nous pouvons donc penser que ces paramètres peuvent apporter de l'information pertinente pour aider à la prédiction de l'intelligibilité en complément d'autres paramètres acoustiques. L'extraction de caractéristiques de l'EMS basée sur de larges bandes de fréquences nous a donc permis de modéliser modéliser en partie le rythme de la parole. La caractérisation automatique de l'EMS nous a permis de prédire efficacement les scores cliniques d'intelligibilité et de sévérité produit par les cliniciens. L'extraction de bandes de fréquence dont les frontières sont similaires d'un locuteur à une autre a en revanche montré une forte dépendance par rapport au débit de parole. En adaptant ces bandes de fréquences aux locuteurs, nous avons pu supprimer cette dépendance au prix d'une dégradation de notre prédiction des scores cliniques. Les paramètres automatiques du rythme basés sur des bandes de fréquence trop larges ne semblent donc pas très adaptés à ce type de tâche. Il serait maintenant pertinent d'extraire de nouveaux paramètres qui se focalisent davantage sur les régularités des niveaux prosodiques des locuteurs de sorte à caractériser plus précisément le rythme de la parole.

5.3 Vers une caractérisation automatique des troubles rythmiques

Dans la partie précédente, nous avons extrait un ensemble de caractéristiques globales de l'EMS en calculant des sommes d'énergie englobant plusieurs niveaux prosodiques. Ces caractéristiques nous ont été utiles pour réaliser des modélisations basées sur un algorithme d'apprentissage automatique. L'un des objectifs de cette thèse est cependant de pouvoir caractériser la parole au travers d'informations sur le rythme. Pour cette tâche, les paramètres que nous avons explorés jusqu'ici ne fournissent pas assez d'information. En effet, mesurer l'énergie sur une bande de fréquences aussi large que [0-4] Hz implique de mélanger des informations à plusieurs niveaux prosodiques comme l'ap, le pw, l'ip alors que nous avons vu dans la section 5.1.1 que l'EMS était capable de fournir des informations avec une bonne granularité. Notre objectif ici est donc de trouver et d'extraire de nouveaux paramètres permettant de pouvoir caractériser plus finement la prosodie des patients.

5.3.1 Raffinement des paramètres automatiques du rythme

Les paramètres extraits précédemment basés sur de larges bandes de fréquences englobent un trop grand nombre d'informations. Ces derniers ne permettent pas une bonne description du rythme de la parole des locuteurs. Afin d'extraire des caractéristiques plus pertinentes, nous devons alors choisir des paramètres qui caractérisent au mieux les pics et leur énergie dans l'EMS. De plus, nous souhaitons pouvoir modéliser le rythme indépendamment du débit de parole qui, comme nous l'avons vu dans la partie 5.2.2, interfère fortement avec nos précédents paramètres. Notre extraction de paramètres s'est donc essentiellement concentrée autour de la caractérisation des pics de l'EMS avec les paramètres suivants :

- La fréquence des deux premiers pics de l'EMS divisée par le débit de parole. Ceci permet d'obtenir la position relative des pics par rapport au débit.
- L'amplitude des deux premiers pics qui indique le degré de régularité rythmique. Plus le pic est haut, plus les durées de l'unité prosodique en question sont similaires. Cette amplitude est normalisée par l'énergie totale du spectre.
- La fréquence et l'amplitude du pic le plus haut au dessus du débit de parole. Ce pic correspond a priori à la régularité syllabique.
- La largeur des trois pics mentionnés ci-dessus. Cette largeur indique l'étendue des durées des unités à laquelle les pics correspondent.
- L'énergie des pics calculée comme la largeur multipliée par l'amplitude.
- Des ratio d'amplitudes et d'énergie entre les deux premiers pics et le pic syllabique.
- Le ratio entre l'énergie de la bande de fréquences [0-ar] Hz (ar = débit articulatoire) et celle de la bande [ar-10] Hz.

Une partie de ces paramètres est illustrée dans la figure 5.13. Nous pensons que l'EMS

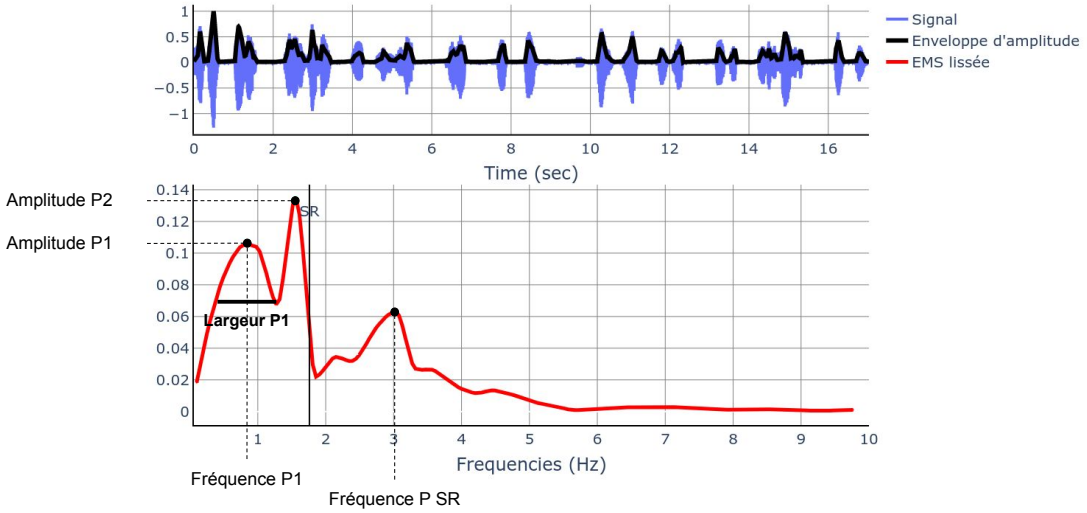

FIGURE 5.13 – Exemple d'extraction des paramètres de l'EMS sur un enregistrement de personne atteinte de cancer VADS (patient n^o 301). Le signal est le même que sur la figure 5.3.

d'un locuteur peut être caractérisée par cet ensemble de paramètres automatiques. De fait, nous espérons pouvoir modéliser le rythme de la parole via leur étude.

5.3.2 Vers une caractérisation des troubles de la prosodie ?

Afin de vérifier la pertinence des paramètres automatiques calculés dans la section précédente (5.3.1), nous pouvons tout d'abord essayer de comparer leurs valeurs avec les scores perceptifs évalués par des cliniciens (décrits dans la section 3.1.5) et les scores prosodiques évalués par un panel de juges naïfs (décrits dans la section 3.2.4). Pour cela, nous avons extrait ces paramètres sur les deux premières phrases de la lecture de texte *"Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon."* pour l'ensemble du corpus VADS. Par la suite, nous avons simplement calculé la corrélation entre les valeurs de nos paramètres et les scores perceptifs. La matrice de corrélation est visible sur la figure 5.14.

Comme nous pouvons le voir, les indices automatiques ne sont que peu corrélés aux scores perceptifs de manière générale. En revanche, trois paramètres ressortent de cette analyse. La fréquence du deuxième pic de l'EMS est légèrement corrélée à plusieurs scores perceptifs avec notamment une corrélation de 0,47 avec la mesure de variabilité rythmique ainsi que -0,43 avec le score clinique de sévérité. La fréquence de ce pic étant normalisée par le débit de parole, ce paramètre nous indique donc la position relative de ce pics par rapport au débit. Ainsi, il semblerait que plus le second pic est proche du débit, plus la prosodie du locuteur paraît dégradée. Sur les exemples que nous avons pu étudier dans la partie 5.1.2, le second pic de l'EMS semble correspondre généralement à la régularité des syntagmes intermédiaires (ip). En revanche, ce niveau n'apparaît pas de façon systématique pour tous les locuteurs.

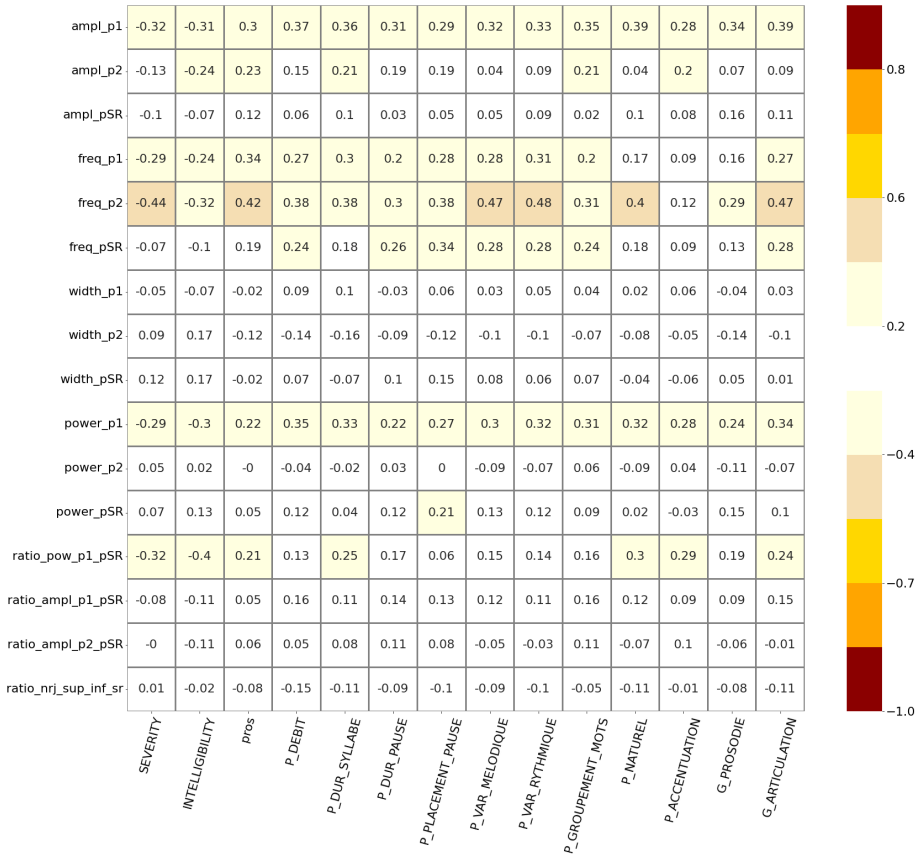

FIGURE 5.14 – Corrélation (Pearson) entre nos paramètres automatiques de l'EMS et les scores perceptifs des dimensions prosodiques décrit dans la partie 3.2.4. 'ampl_px' correspond à l'amplitude maximale du pic numéro 'x' (SR désigne le premier après le débit de parole), 'freq_px' est la fréquence du pic divisé par le débit de parole, 'width_px' est la largeur du pic, 'power_px' correspond à la largeur multiplié par l'amplitude. Le dernier paramètre 'ratio_nrj_sup_inf_SR' désigne le ratio entre l'énergie de les bandes de fréquences [0-SR] et [SR-10] Hz. Les corrélations en dessous de 0.2 ne sont pas affichées.

Il est donc possible qu'il représente le niveau du syntagme accentuel sur certains enregistrements.

Les deux autres paramètres automatiques pertinents sont l'amplitude (ampl_p1) et l'énergie (power_p1) du premier pic de l'EMS. Ces indices sont respectivement corrélés à hauteur de 0,39 et 0,32 avec la variabilité rythmique ou encore 0,31 et 0,4 avec l'intelligibilité. Selon nos observations, ce pic correspond généralement à la régularité du syntagme intonatif (IP) mais peut représenter également l'ip si l'IP n'est pas produite de façon régulière par les locuteurs. Il semblerait donc qu'avoir une grande régularité rythmique des syntagmes intonatifs dégraderait la perception de la prosodie. Bien que contre-intuitif, en réalité, ceci va dans le sens de l'hypothèse que nous avions émise dans la section 5.1.1. En effet, nous avions remarqué que les personnes fortement atteintes par leur cancer compensent leurs troubles articulatoires par une structure prosodique très régulière.

Notre hypothèse semble donc se vérifier sur ce corpus. Pour aller plus loin, nous avons essayé de modéliser le rythme des patients VADS en projetant les locuteurs dans un espace en deux dimensions basées sur nos indices automatiques. Pour cela nous avons essayé de nombreuses combinaisons de paramètres automatiques. La modélisation la plus pertinente selon nous provient des deux paramètres les plus corrélés avec les annotations perceptives : la fréquence normalisée du deuxième pic de l'EMS et la puissance du premier pic. Nous avons donc projeté notre sous-ensemble de 20 locuteurs (10 contrôles et 10 VADS) sur ces deux dimensions afin de pouvoir comparer ces dimensions à la catégorisation libre réalisée par Corine Astésano sur ces mêmes locuteurs. La méthodologie de cette catégorisation a été décrite dans la partie 3.2.3 et est décrite dans la figure 3.12 (p.72). La projection via nos paramètres est visible sur la figure 5.15.

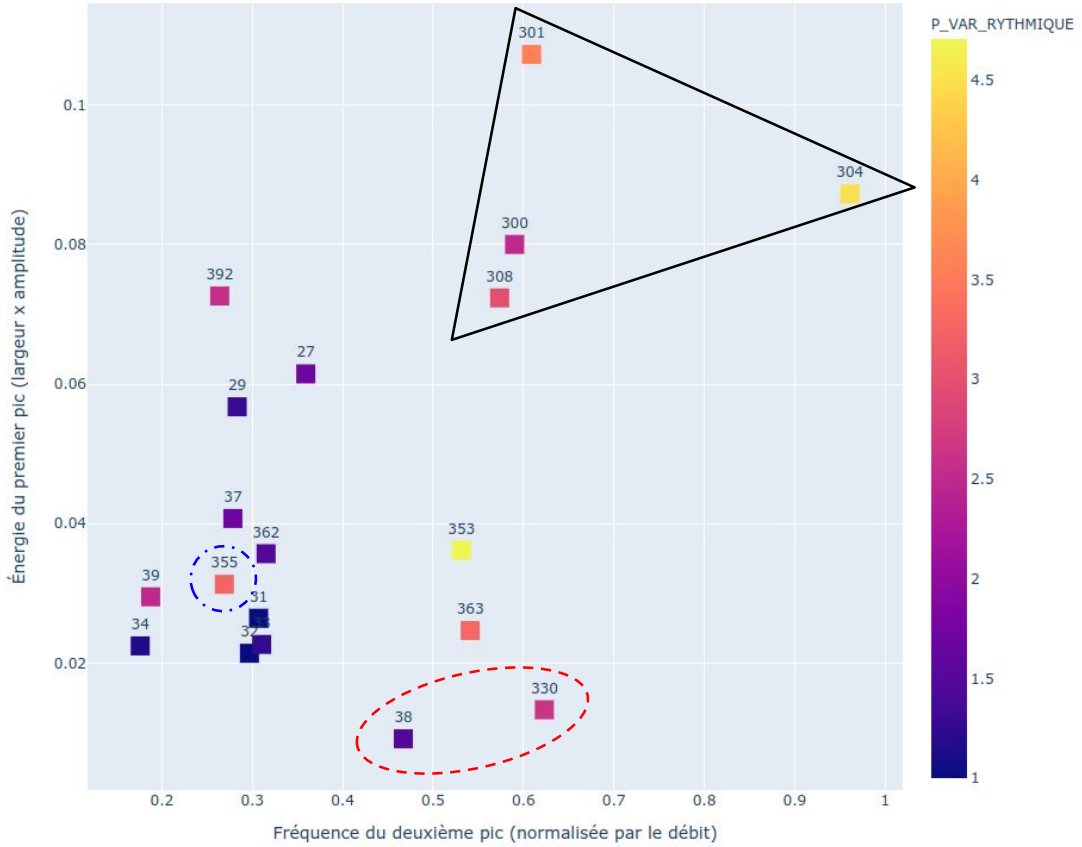

FIGURE 5.15 – Représentation des 20 locuteurs (10 témoins, 10 VADS) que nous avons sélectionnés dans la partie 3.2.3. L'axe des abscisses correspond à la fréquence du deuxième pic de l'EMS. L'axe des ordonnées correspond à l'énergie du premier pic de l'EMS (amplitude maximale du pic multipliée par sa largeur). Les sujets témoins sont indiqués par un numéro inférieur à 100, les sujets VADS ont un numéro supérieur à 300. La couleur des points est déterminée par les scores perceptifs de variabilité rythmique (un score élevé correspond à une grosse variabilité).

À partir de cette projection en deux dimensions, nous pouvons constater un regroupement en partie similaire à notre catégorisation libre, avec en haut à droite (dans le triangle noir), les locuteurs 300, 301, 304 et 308 qui sont proches dans la catégorisation libre et dans notre représentation. De même, au niveau des sujets contrôles (n^o inférieures à 100), les sujets 38 et 27 sont éloignés du groupe de sujets contrôles, qui ont été jugés perceptivement moins fluents que les autres. Le sujet 38 est d'ailleurs placé proche du patient 330 (cercle rouge en bas de la figure) ce qui est en accord avec la perception, puisque ces deux locuteurs ont été jugés comme ayant de bonnes cibles articulatoires mais une fluence dégradée. En revanche, certains résultats sont plus difficiles à interpréter, comme avec le patient n^o 355 qui est placé au centre des sujets contrôles (rond bleu sur la figure) alors qu'il est plus proche des patients les plus touchés dans la catégorisation. Pour essayer de comprendre cela, nous pouvons regarder l'EMS de ce locuteur qui se trouve dans la figure 5.16.

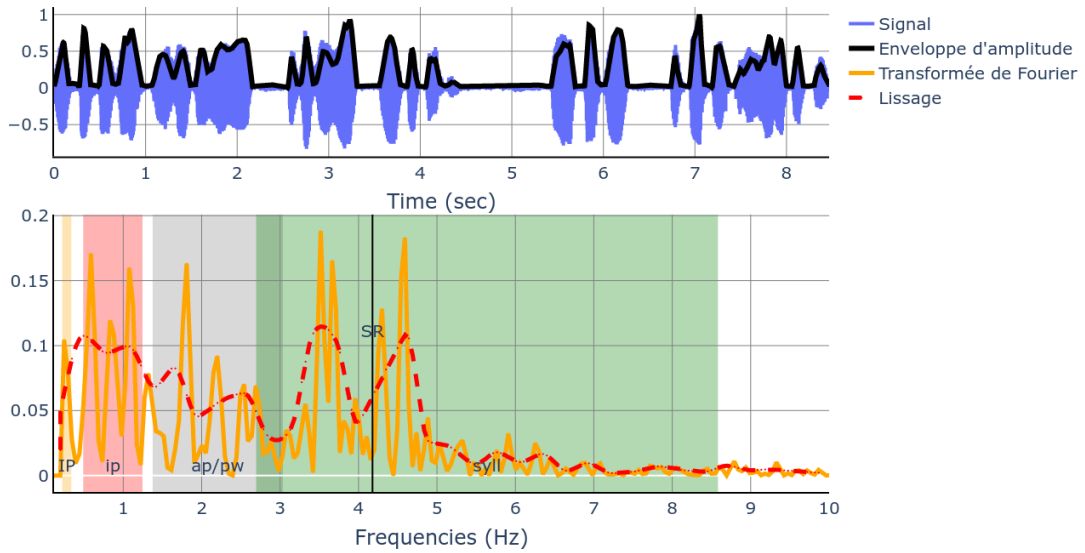

FIGURE 5.16 – EMS du locuteur cancer VADS n^o 355 (sévérité = 2,66) sur les phrases "*Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon*". Les intervalles des niveaux prosodiques sont indiqués en couleur : rouge pour l'ip, gris pour l'ap (ici environ équivalent au pw) et vert pour la syllabe.

À la vue de cet EMS, il est difficile de savoir si le locuteur présente des troubles rythmiques car il possède des pics de régularités à plusieurs niveaux et des pics syllabiques liés à la dimension articulatoire bien présents. Après avoir réécouté cet enregistrement, nous avons pu constater que ce locuteur présente une forte nasalité. L'accès au sens est donc fortement diminué au niveau perceptif. En revanche, du point de vue prosodique, la fluence et la structuration rythmique de cet énoncé ne nous semble pas fortement dégradée. Nous pensons donc que le placement de ce patient dans la catégorisation libre (figure 3.12 p.72) a été influencé par la sévérité de sa pathologie. Il est tout de même important de rappeler que la catégorisation libre repose

sur l'étude de l'ensemble du texte tandis que notre représentation ne se concentre que sur les deux premières phrases. Une future étude devra donc se concentrer sur l'inclusion de l'ensemble du texte dans cette modélisation.

Maintenant que nous avons validé notre représentation sur notre extrait du corpus VADS, nous pouvons réaliser la même procédure sur l'ensemble des locuteurs. Le résultat obtenu est visible sur la figure 5.17. Nous pouvons remarquer ici que les

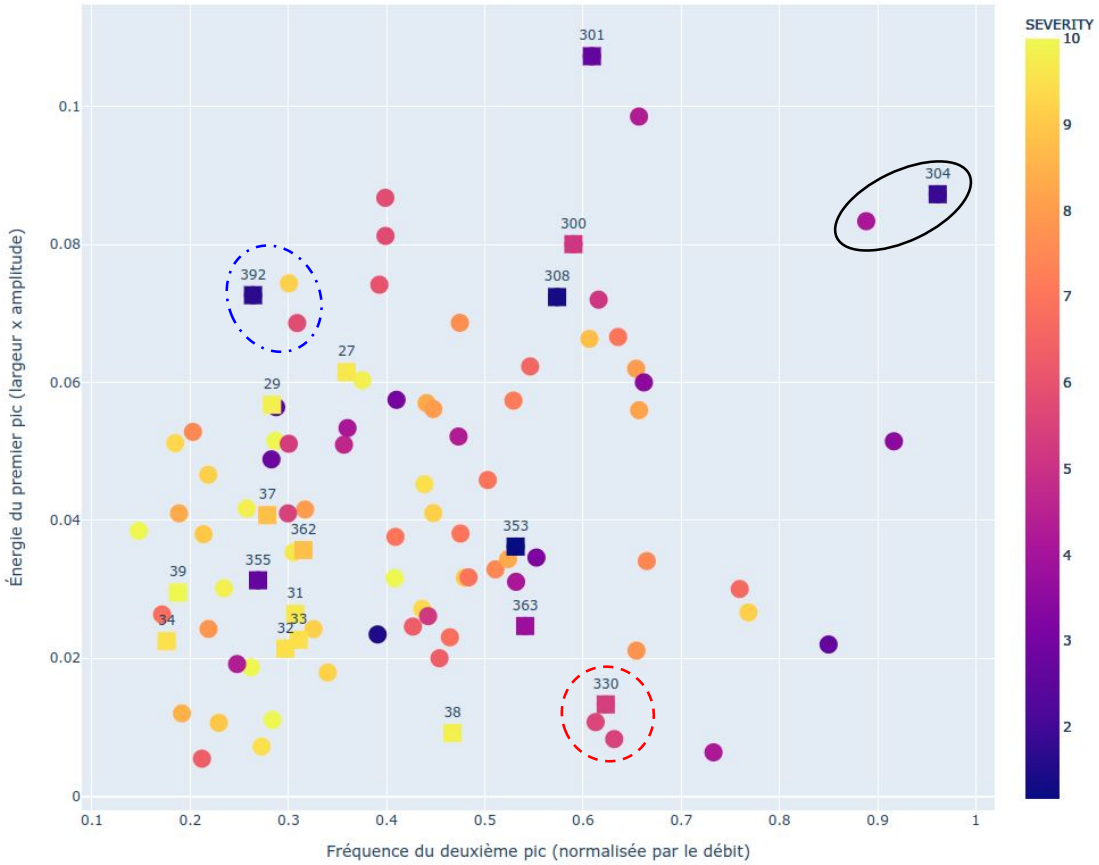

FIGURE 5.17 – Représentation de l'ensemble des locuteurs du corpus VADS. L'axe des abscisses correspond à la fréquence du deuxième pic de l'EMS. L'axe des ordonnées correspond à l'énergie du premier pic de l'EMS (amplitude maximale du pic multipliée par sa largeur). La couleur des points est déterminée par les scores cliniques de sévérité (un score faible correspond à une maladie très sévère). Les marqueurs carrés correspondent aux personnes issues de notre sélection de 20 locuteurs.

personnes avec une sévérité normale (locuteurs jaunes / orange) ont tendance à être regroupées dans la partie inférieure gauche de la figure, ce qui confirme notre observation sur notre sélection de 20 locuteurs. De même, nous avons écouté des enregistrements de sujets proches les uns des autres dans la projection pour vérifier la cohérence de ses dimensions. Ainsi, le patient proche du locuteur n° 304 (dans l'ellipse noire) présente le même profil avec une segmentation de la parole en unités très courtes et des pauses entre chaque mots prosodiques. En bas de notre représentation (dans le cercle rouge), le patient 330 et les deux locuteurs proches de lui ont également certaines

caractéristiques prosodiques similaires tout en ayant des caractéristiques acoustiques différentes. Le patient 330 possède une voix très éraillée avec des problèmes de phonations tandis que les deux autres ne présentent pas cette particularité. Au niveau de leurs structures rythmiques, nous retrouvons en revanche une irrégularité des durées des grandes unités prosodiques probablement due à des troubles dans leur lecture. À gauche de notre projection en revanche, les deux locuteurs proches du n° 392 (ellipse bleue) ne semblent pas présenter les mêmes caractéristiques prosodiques. Le 392 présente de mauvaises cibles articulatoires contrairement aux deux autres et nous n'avons pas su repérer leurs similarités prosodiques.

La modélisation automatique de la structuration rythmique que nous proposons n'est donc pas parfaite mais permet tout de même de catégoriser certains types de paroles. Nous distinguons notamment les stratégies de segmentations des énoncés des locuteurs. Les personnes produisant les segmentations en ip / IP les plus régulières sont ainsi regroupées entre elles. Nous avons également pu remarquer que les personnes sans troubles apparents de la parole se regroupent davantage dans la partie inférieure droite de notre projection. Cette zone correspond aux personnes dont le premier pic de l'EMS est faible, indiquant une mauvaise régularité des durées des syntagmes intermédiaires ou intonatifs. Du côté opposé, nous retrouvons dans la partie supérieure droite des locuteurs dont les ip et/ou IP sont très régulières. De même, le second pic de leur EMS est proche de leur débit de parole (ratio de la fréquence du pic sur le débit de parole proche de 1) ce qui signifie que ce second pic correspond à une unité proche de la syllabe comme le mot prosodique. Les personnes à droite de notre visualisation produisent donc davantage de groupes courts (pw) et les assemblent dans des groupes de durée très régulière.

Notre modélisation ne devrait en revanche pas permettre de différencier efficacement les personnes en fonction de leurs troubles articulatoires étant donné que nous nous focalisons seulement sur les unités rythmiques de niveaux supérieurs. Nous supposons néanmoins que les déficits lourds d'articulations se retrouvent dans notre axe des ordonnées qui mesure l'énergie du premier pic de l'EMS. En effet, les locuteurs situés dans la partie haute de notre figure sont principalement des personnes atteintes de troubles majeurs de l'articulation. Plusieurs hypothèses sont alors possibles. Premièrement, il est possible que plus les cibles articulatoires des patients sont mauvaises, plus leur énergie dans les zones de fréquences supérieures au débit est faible. Cette baisse d'énergie dans la partie droite de l'EMS se répercute alors par une hausse globale de l'énergie de la partie gauche et donc une hausse du premier pic. En plus de cette hypothèse concernant des particularités de calculs de l'EMS, nous pouvons également supposer que ce phénomène appuie l'hypothèse que nous avons émise dans la section 5.1.1.

Selon nous, il est possible que les personnes atteintes de forts troubles articulatoires compensent leurs déficits en produisant une structure prosodique très régulière de leur parole. Cette régularité se traduirait alors par une forte intensité du premier pic de l'EMS. Il n'est donc à l'heure actuelle pas possible d'affirmer avec certitude que notre hypothèse s'avère vraie, mais il est nécessaire de poursuivre les travaux

dans ce domaine. Cela pourra alors être réalisé via l'étude de plus grands corpus de parole pathologique, sur des tâches de production de parole spontanée qui permettront d'évaluer la prosodie des patients dans des conditions plus proches de leur quotidien.

5.4 Conclusion de chapitre

Dans ce chapitre, nous avons décidé d'utiliser le spectre de modulations d'amplitude (EMS) pour pouvoir étudier de façon automatique le rythme de la parole pathologique. Pour cela, nous avons utilisé les deux corpus de paroles pathologiques décrits dans la section 3.1. Dans un premier temps, nous avons réalisé une analyse qualitative de plusieurs locuteurs parmi des sujets contrôles, des personnes atteintes de cancer VADS et des patients atteints de la maladie de Parkinson. Nous avons alors pu constater que l'EMS d'un sujet sain présente généralement des pics de régularité sur l'ensemble de la bande de fréquence [0-10] Hz avec des pics de hauteurs variées. Les patients MDP, quant à eux présentent relativement peu de troubles de la prosodie pour la majorité des patients. En revanche, nous avons remarqué des troubles rythmiques chez certains locuteurs MDP qui produisent des segments irréguliers de paroles avec des placements et des durées de pauses inhabituels. Cela se traduit sur l'EMS de ces patients par une absence majeure de pics de régularité à plusieurs niveaux. Enfin, chez les sujets cancers VADS, nous avons pu observer la présence de troubles articulatoires majeurs et une baisse du débit de parole. La structuration prosodique de ces patients en revanche n'est pas désorganisée. En effet, leurs énoncés sont structurés par des groupes prosodiques courts (pw, ap) de durées très régulières. Ces petits groupes se regroupent alors en unités plus larges tout aussi régulières. Au niveau de leur EMS, nous retrouvons donc une très faible énergie dans les bandes de fréquences supérieures à 5 Hz étant donné leurs forts troubles articulatoires, mais en contrepartie, l'énergie dans les zones de fréquences plus faibles est élevée. Cette observation nous a mené à émettre l'hypothèse selon laquelle les personnes atteintes de troubles articulatoires sévères compensent leurs déficits via une structure rythmique très régulière. Afin de nous assurer de la pertinence de notre modélisation via l'EMS, nous avons utilisé les annotations prosodiques détaillées dans la section 3.2.2. Ces annotations décrivent les unités prosodiques des énoncés de 20 locuteurs (10 contrôles et 10 sujets VADS). En superposant les durées moyennes de ces unités avec l'EMS, nous avons pu montrer que les pics du spectre correspondent effectivement aux différentes unités prosodiques (IP, ip, ap, pw, syllabes).

Par la suite, nous avons voulu caractériser automatiquement les EMS des locuteurs en extrayant des paramètres automatiques simples qui englobent des informations à plusieurs niveaux. Nous avons donc extrait des paramètres basés sur l'amplitude et la fréquence des pics principaux de l'EMS ainsi que l'énergie dans des bandes de fréquences larges : [0-4] Hz et [4-10] Hz (Liss et collab., 2010). Une fois ces caractéristiques extraites, nous avons voulu essayer de les utiliser afin de réaliser une prédiction automatique de l'intelligibilité et la sévérité clinique (décrites dans la section

3.1.5). L'objectif de cette étude était donc de voir si le rythme joue ou non un rôle dans l'amélioration de l'intelligibilité des patients. Nous avons alors utilisé l'algorithme SVR (*Support Vector Regression*) d'apprentissage automatique. Les résultats ont montré que nous arrivions à prédire efficacement l'intelligibilité avec une corrélation de 0,73 entre nos prédictions et l'intelligibilité de référence pour le corpus VADS. Sur le corpus Parkinson, cette corrélation est de 0,68. Le paramètre automatique le plus utilisé par notre modélisation automatique pour prédire les scores cliniques est le rapport entre l'énergie dans la bande de fréquences [0-4] Hz et celle dans la bande [4-10] Hz.

Nous avons cependant remis en question nos résultats en remarquant que nos paramètres les plus utilisés étaient fortement corrélés au débit de parole. La normalisation de ces paramètres par le débit articulatoire (nombre de syllabes divisé par la durée de phonation) nous a mené à remplacer nos bandes d'énergies fixes par des bandes adaptées à chaque locuteur en fonction de son débit articulatoire. Nous avons donc extrait l'énergie dans les bandes [0-AR] Hz et [AR-10] Hz (AR : débit articulatoire). En utilisant ces nouveaux paramètres normalisés, la qualité de notre prédiction diminue fortement. Ce résultat suggère plusieurs choses. Tout d'abord que les annotations cliniques de sévérité et intelligibilité sont fortement corrélés au débit de parole, mais également que nos paramètres automatiques ne sont pas idéaux pour modéliser précisément le rythme de la parole.

Notre modélisation n'étant pas idéale pour décrire avec précision la structuration rythmique des locuteurs, nous avons essayé d'extraire de nouveaux paramètres automatiques afin d'extraire des informations plus fines de l'EMS. Pour cela, nous nous sommes concentrés davantage sur les pics de l'EMS en lieu et place des bandes d'énergies trop larges. Ainsi, les nouveaux paramètres descriptifs de l'EMS concernent l'amplitude, la largeur et la fréquence des principaux pics de l'EMS. Afin de ne pas être impactées par le débit de parole, nos mesures de fréquences des pics ont été normalisées par la valeur estimée du débit. Afin de valider la pertinence de ces caractéristiques, nous avons mesuré leurs corrélations avec les scores issus de l'analyse prosodique perceptive effectuée dans la section 3.2.4. Deux paramètres principaux ressortent de cette analyse. Le premier est la fréquence (normalisée par le débit) du second pic de l'EMS avec une corrélation de 0,47 avec le score perceptif de variabilité rythmique. Le second concerne l'énergie contenue dans le premier pic de l'EMS qui est légèrement corrélée avec la majorité des scores perceptifs et cliniques.

En projetant l'ensemble des sujets atteints de cancers VADS dans un plan composé de ces deux dimensions, nous avons pu analyser les placements des différents locuteurs en fonction de leurs propriétés rythmiques. Nous avons alors comparé cette visualisation à la catégorisation libre de la prosodie des locuteurs qui a été décrite dans la section 3.2.3. Une grande partie des regroupements faits dans cette catégorisation libre a été retrouvé dans notre modélisation du rythme. Ainsi, la fréquence du second pic de l'EMS indique selon nous à quel niveau prosodique correspond ce pic. Une fréquence élevée signifie que ce pic correspond probablement à l'ap ou au pw et donc que soit l'ip ou l'IP n'est pas représentée. Cette forte fréquence indiquerait alors quel niveau est prédominant dans la parole du locuteur. L'énergie du premier

pic quant à lui indique le degré de régularité du niveau prosodique le plus large qui est généralement l'IP (parfois l'ip). Une forte énergie indiquerait alors que les IP (ou ip) du locuteur sont de durées très similaires. Nous pensons également qu'une forte énergie de ce pic peut signifier une baisse d'énergie dans les bandes de fréquences supérieures (syllabes/phonèmes). Cette dernière possibilité pourrait donc aller dans le sens de notre hypothèse selon laquelle les patients atteints de troubles sévère de l'articulation ont tendance à produire un rythme plus régulier de leur parole.

À ce jour, il est encore difficile d'affirmer que les résultats fournis par l'EMS pour caractériser le rythme de la parole sont parfaitement fiables. Nous n'avons en effet testé cette méthodologie que sur une faible quantité de données. De plus, il est important de rappeler que notre étude porte sur la lecture de texte. Ceci nous permet d'obtenir des résultats comparables d'un locuteur à un autre grâce à un texte commun. Cependant, il est certain que la lecture de texte ne représente pas la réalité prosodique de la parole spontanée. Par la suite, il sera donc nécessaire de continuer à explorer ce type de modélisations au travers d'études sur de plus grands corpus de parole pathologique pour des tâches de production de parole spontanée qui permettront d'évaluer leurs déficits prosodiques et ainsi de mesurer l'impact que peuvent avoir ces déficits sur la qualité de vie de ces patients.

Conclusions et perspectives

Conclusion générale

L'objectif principal de notre étude était de modéliser automatiquement le rythme de la parole de sorte à pouvoir évaluer les troubles de la structuration prosodique dans le cadre de la parole pathologique. Ce travail à l'interface entre la linguistique et le traitement du signal nous a demandé de passer par plusieurs étapes successives que sont : circonscrire le concept de rythme de la parole, trouver des modélisations automatiques du rythme utilisées dans la littérature, implémenter et tester ces modélisations sur de la parole continue et enfin les appliquer à la parole pathologique.

Dans le chapitre 1, nous nous sommes attelés à poser les fondements théoriques du rythme sur lesquels nous nous sommes reposés tout au long de ce travail. Nous avons exposé la vision de la théorie métrique, développée par (Liberman, 1975; Liberman et Prince, 1977) et adaptée au français par (Di Cristo, 2000). Dans cette vision, l'accentuation est placée au centre de la structuration prosodique. L'alternance des syllabes (accentuées ou non) est formalisée au travers d'une organisation hiérarchique et de la notion de poids métrique qui pondère les différentes unités prosodiques. Nous avons alors présenté les différentes unités que nous considérons : la syllabe, le mot prosodique (pw), le syntagme accentuel (ap), le syntagme intermédiaires (ip) et le syntagme intonatif (IP). Lors de la production de parole, ces niveaux ne sont cependant pas strictement réguliers. En effet, le rythme est considéré comme la manifestation de surface de la métrique sous-jacente. Le rythme est cependant également dépendant de nombreuses contraintes physiologiques, syntaxiques et pragmatiques qui induisent une certaine variabilité. Nous avons néanmoins émis l'hypothèse que la régularité accentuelle autour de 550 ms pouvait se retrouver dans nos corpus de parole. Il s'agit d'un rythme observé dans plusieurs langues par (Fant et collab., 1991), et qui a été mis en évidence pour le français par (Astésano, 2001). Nous pensons que cette fenêtre temporelle correspondant au mot prosodique pw est particulièrement pertinente pour la parole pathologique, puisque les problèmes articulatoires des patients cancers peuvent les conduire à adopter une stratégie de planification de la parole au niveau du mot prosodique. Il n'est cependant pas évident d'étudier automatiquement le rythme de la parole, car les indices acoustiques de l'accentuation ne sont pas toujours en adéquation avec la perception des proéminences (Astésano, 2017; Astésano, 2019). La

modélisation automatique du rythme est donc un sujet complexe qui doit tenir compte d'un grand nombre de paramètres.

Dans le chapitre 2, nous avons exposé une évolution des propositions de modélisations du rythme des années 1990 à aujourd'hui, à travers l'analyse des durées inter-vocaliques et inter-consonantiques (Ramus et collab., 1999; Ling et collab., 2000; Grabe et Low, 2002; Dellwo, 2006). L'analyse de ces durées a montré des résultats encourageants dans le cadre de l'identification des langues en utilisant des méthodes totalement automatiques (Farinas, 2002; Pellegrino et collab., 2002). Mais ces mesures ne fournissent pas une modélisation satisfaisante du rythme de la parole par rapport à notre vision du rythme exposée dans le premier chapitre. C'est la raison pour laquelle nous nous sommes intéressés à d'autres domaines, comme celui de la musique, où le rythme est modélisé en tenant compte des niveaux de structuration hiérarchiques. En musicologie, le rythme principal perçu est appelé le tempo et plusieurs méthodologies ont été développées de façon à extraire cette valeur automatiquement. Parmi ces techniques, nous nous sommes particulièrement intéressés à une sous-partie d'entre elles. Elles se basent sur l'étude de régularités rythmiques (Desain, 1992; Miguel Alonso et Richard, 2004). Pour cela, une transformée de Fourier est appliquée à une segmentation en "onsets" des battements principaux. Cette représentation permet alors d'extraire le tempo principal d'une musique. Cette méthodologie a été appliquée à la parole afin d'en analyser le rythme au travers d'une modélisation appelée le Tempogramme (Le Coz, 2014). Cette méthode permet théoriquement d'explorer les régularités rythmiques à différents niveaux et non plus seulement au niveau syllabique. Dans cette même optique, plusieurs travaux ont été réalisés en se basant sur un principe similaire mais n'utilisant pas de segmentation du signal : les spectres de modulations (EMS) (Tilsen et Johnson, 2008; Sheft et collab., 2008, 2012; Varnet et collab., 2017). Ces spectres de modulations ont permis de modéliser le rythme de façon à correspondre au mieux à notre vision linguistique, i.e. modéliser la hiérarchie de la parole en utilisant les variations d'amplitude et de fréquences du signal. L'utilisation de l'EMS a de fait été utilisée dans plusieurs domaines dont l'étude de la parole pathologique (Liss et collab., 2010). Nous avons voulu appliquer certains de ces algorithmes sur les corpus de parole que nous avons détaillés dans le chapitre 3.

Un ensemble de trois corpus ont été exposés dans ce chapitre avec notamment un corpus restreint de slam annoté manuellement au niveau du rythme (Simon, 2020) que nous avons utilisé afin de valider la pertinence des modèles décrits dans la section 2. Nous avons ensuite présenté nos deux corpus de parole pathologique issus du projet ANR RUGBI. Le premier corpus concerne de la parole de personnes atteintes de cancers des Voies Aéro-Digestives Supérieures (VADS) constitué de 87 patients et 26 sujets contrôles (Woisard et collab., 2021). Le second corpus contient des enregistrements de parole de personnes atteintes de la maladie de Parkinson (MDP) avec au total 205 patients et 111 sujets contrôles (Ghio et collab., 2012). Ces corpus de parole pathologique sont constitués d'un ensemble de tâches de production de parole diverses. Nous

avons choisi pour nos travaux de nous focaliser sur la tâche de lecture du texte de *la chèvre de Monsieur Seguin*. Cette tâche n'est pas idéale afin de rendre compte du rythme de la parole dans des conditions réelles de parole spontanée. Cependant, elle nous permet de pouvoir obtenir des résultats comparables d'un locuteur à un autre. Au delà des enregistrements, nos corpus de parole pathologique sont constitués d'un ensemble d'annotations perceptives de la prosodie que nous avons récupéré ou généré. Parmi elles, plusieurs annotations cliniques ont été mises à notre disposition. Nous avons notamment des scores de sévérité et d'intelligibilité (de 0 à 10) évalués par un ensemble de six professionnels de santé. L'intelligibilité fait référence à la qualité de reconstruction d'un énoncé par un auditeur au niveau acoustico-phonétique (Pommée et collab., 2022). La sévérité en revanche fait davantage référence à la mesure de l'impact d'une pathologie sur la production orale de façon générale (Woisard et Lepage, 2010).

En plus des scores cliniques, nous avons pu utiliser plusieurs types d'annotations prosodiques sur la tâche de lecture. Une partie de ces annotations s'est concentrée sur l'étude d'un extrait du corpus VADS (10 patients et 10 témoins). Nous avons ainsi généré pour ces 20 personnes des annotations rythmiques indiquant les durées des différents niveaux prosodiques qu'ils produisent (syllabes, ap, pw, ip, IP). De plus, cet extrait de corpus a fait l'objet d'une catégorisation libre du rythme visible sur la figure 3.12 (p. 72). Cette catégorisation a été réalisée de sorte à pouvoir obtenir une notion de distance entre plusieurs locuteurs afin de pouvoir évaluer la pertinence des modélisations automatiques du rythme. En plus de la constitution de ces annotations, nous avons pu récupérer des scores perceptifs d'évaluation de la prosodie jugés par un panel de huit personnes naïves. Ces scores nous ont permis de mieux comprendre l'impact de la prosodie sur les scores d'intelligibilité et de sévérité produits par les médecins.

Dans le chapitre 4, nous nous sommes penchés sur la sélection des modélisations du rythme de la parole. Pour cela, nous avons éprouvé plusieurs méthodologies sur notre corpus de slam dans le but de vérifier que les modélisations rendent compte ou non des régularités rythmiques qui émergent à plusieurs niveaux. Le slam étant un style de parole très rythmé, nous espérions pouvoir éprouver nos méthodologies. La première méthode testée a été le tempogramme (Grosche et collab., 2010; Le Coz, 2014). Malgré de nombreuses expérimentations dans le but d'adapter cette méthode à la parole, nous n'avons pas réussi à rendre compte des régularités rythmiques des niveaux prosodiques autres que celui de la syllabe. Nous avons donc abandonné le tempogramme au profit du spectre de modulations d'amplitude et de fréquence. Ces modélisations du rythme de la parole nous ont alors fourni une représentation visuelle cohérente de la régularité des différents niveaux prosodiques mis en jeu dans les textes de slam (figure 4.13 p.91). Cette méthodologie nous a également permis d'estimer les durées correspondantes à ces niveaux sur l'ensemble des unités allant de 100 ms à plusieurs secondes (en fonction de la durée de l'énoncé). Néanmoins, le spectre de modulations de fréquence a montré quelques limites. En ne se basant que sur l'étude

des variations de l'intonation via la modélisation Momel (Hirst et Espesser, 1993), nous avons contraint notre modèle à se focaliser sur les régularités rythmiques de haut niveau tel que l'IP qui est le domaine de variation de l'intonation. Ceci entraînant alors un manque de précision sur les régularités syllabiques. Le spectre de modulations d'amplitude apportant une information au moins aussi complète que celui de fréquence, nous avons donc décidé de l'utiliser pour modéliser automatiquement le rythme de la parole pathologique dans le chapitre 5.

Une fois notre méthodologie éprouvée sur notre corpus de slam, nous l'avons appliquée à la parole pathologique. Nous avons alors pu comparer l'EMS de plusieurs locuteurs en nous concentrant sur les particularités rythmiques caractéristiques des populations contrôle, MDP et cancer VADS. Chez les personnes saines ne présentant pas de troubles de la parole, nous observons un EMS composé de nombreux pics répartis équitablement sur la bande de fréquence [0-10] Hz. Ceci montre que ces personnes produisent une parole dont le rythme est présent à plusieurs niveaux prosodiques simultanément (voir figure 5.2 p.106) comme le décrit le modèle théorique des oscillateurs couplés (décrit dans la section 1.2.2). Les patients atteints de la MDP ont, pour certains, quelques troubles prosodiques concernant leur fluence. Ces patients produisent parfois des énoncés mal structurés décomposés en longues pauses de durées aléatoires. Chez ces locuteurs, l'EMS produit très peu de pics avec généralement un pic correspondant à l'IP (comme sur la figure 5.1 p.103). Les déficits prosodiques des patients MDP nous semblent relativement difficiles à étudier du point de vue du rythme étant donné qu'une grande majorité des patients ne présentent pas de troubles majeurs de la prosodie. Du côté des cancers VADS en revanche, ce type de troubles est davantage présent chez les patients. Nous avons donc repéré des patients avec de forts troubles articulatoires dont la prosodie est particulière. Par exemple, les figures 5.3 et 5.4 (p.107) nous montrent les EMS de patients avec des structurations rythmiques basées respectivement sur le pw et sur l'ip. Les troubles articulatoires de ces patients les forcent à produire des énoncés longs de façon à pouvoir se concentrer sur la production de leur cibles articulatoires. Ainsi, cette contrainte implique une structuration en unités prosodiques plus ou moins courtes dont les durées avoisinent les 500 ms (cf. 1.2.3).

À partir de ces analyses, nous avons choisi d'extraire quelques paramètres de façon automatique sur l'EMS. L'objectif de cette extraction étant de pouvoir caractériser les particularités rythmiques d'un patient. Pour cela, nous nous sommes dans un premier temps basé sur les travaux de Liss et collab. (2010). Les paramètres sont basés sur l'extraction d'énergies dans des bandes de fréquences larges ([0-4] et [4-10] Hz par exemple) englobant plusieurs niveaux prosodiques. Afin de tester ces paramètres, nous avons réalisé une modélisation automatique de l'intelligibilité via une technique d'apprentissage automatique supervisée (Drucker et collab., 1997). Nous avons alors pu produire une prédiction fiable des scores cliniques (voir figures 5.7 et 5.8 p.113-114) tant sur le corpus Parkinson (erreur moyenne absolue pour l'intelligibilité = 0.27) que cancer VADS (erreur moyenne absolue pour l'intelligibilité = 1.01). Ces scores sont

cependant à considérer avec précaution étant donné que la majorité des locuteurs (voire la grande majorité pour le corpus MDP) ont une intelligibilité supérieure à 8. Une fois notre modèle en place, nous nous sommes rendu compte que l'écrasante majorité de notre prédiction était calculée grâce au ratio entre l'énergie de l'EMS dans la bande [0-4] Hz et celle entre [4-10] Hz. Cependant, ce paramètre étant fortement corrélé au débit de parole, nous avons dû neutraliser la relation de dépendance entre ces variables. Cette neutralisation avait pour but de nous concentrer uniquement sur des informations rythmiques indépendantes du débit de parole. En neutralisant l'effet du débit, en choisissant les bandes [0-ar] et [ar-10] Hz (ar : débit articulatoire), nous avons pu diminuer la dépendance entre notre ratio et le débit de parole. Néanmoins, en réitérant la modélisation des scores cliniques avec ces nouveaux paramètres, la qualité de nos prédictions a chuté fortement (voir figures 5.11 et 5.12 p.117-118). Le ratio entre [0-ar] et [ar-10] n'est pas forcément une information pertinente pour caractériser l'intelligibilité et la sévérité de la parole de patients atteints par les cancers VADS. Il serait intéressant de chercher à mieux caractériser la prosodie de ces personnes pour le mettre en rapport avec la compréhensibilité du message.

À la suite de nos expérimentations, nous avons tout de même voulu essayer de modéliser le rythme de la parole pathologique en nous abstenant au maximum d'utiliser le débit de parole. Ainsi, nous avons caractérisé l'EMS au travers de nouveaux paramètres automatiques basés sur les caractéristiques de ses pics, en espérant pouvoir modéliser les régularités des niveaux prosodiques. Ces nouveaux indices basés sur des mesures d'amplitudes, de fréquences (normalisées par le débit) et d'énergies des pics ont alors pu être comparés aux annotations prosodiques présentés dans la section 3.2.4 afin de pouvoir obtenir une bonne interprétabilité des indices automatiques. Nous avons alors mesuré la corrélation des indices avec les scores de l'analyse prosodique perceptive et identifié deux paramètres principaux. La fréquence du second pic de l'EMS qui est corrélée à 0,47 avec le score de variabilité rythmique. Le second indice est l'énergie du premier pic de l'EMS qui est légèrement corrélé avec la plupart des scores perceptifs et cliniques. Dans l'optique d'une caractérisation automatique des troubles rythmiques de la parole pathologique, nous avons projeté l'ensemble des locuteurs du corpus cancers VADS dans un plan formé par ces deux indices automatiques (cette projection est visible sur la figure 5.17). Nous avons alors réussi à trouver une correspondance entre certains groupes issus de la catégorisation libre de la prosodie décrite dans la section 3.12. Nous supposons que la fréquence normalisée du second pic peut indiquer l'unité prosodique auquel le pic correspond. Une fréquence élevée correspond à une unité courte (pw / ap). L'énergie du premier pic concerne davantage la régularité du niveau le plus large comme l'IP ou l'ip. Nous pensons que cette énergie peut également refléter une baisse d'énergie dans les bandes de fréquences supérieures chez les patients atteints de troubles de l'articulation sévères.

Les résultats de nos travaux ont donc permis de dégager plusieurs caractéristiques rythmiques des pathologies étudiées. En particulier chez les patients atteints de cancers VADS où nous avons pu localiser des stratégies de structuration rythmiques plus

petites et plus régulières chez les patients atteints de troubles articulatoires sévères. La taille et la régularité des constituants prosodiques seraient alors proportionnels à la sévérité des déficits articulatoires. De nombreux travaux sont encore à effectuer sur la pertinence de l'EMS pour caractériser le rythme. En effet, nos travaux ne se sont concentrés que sur quelques phrases isolées de lecture de texte. La prochaine étape sera donc d'appliquer nos travaux à la parole spontanée dans le but de nous rapprocher des conditions réelles de production de la parole. Nous pourrons ainsi caractériser les stratégies prosodiques utilisées par les patients dans leur quotidien.

Perspectives

Ce travail de doctorat amène de nombreuses pistes de recherche, et ouvre de nombreuses perspectives. Certaines d'entre elles pourraient même représenter des défits à même d'alimenter plusieurs doctorats !

Pistes d'amélioration du spectre de modulations

Nous avons pu voir dans les chapitres 4 et 5 que le spectre de modulations d'amplitude lissé permettait de modéliser efficacement les régularités rythmiques de la parole continue. Cependant, cette modélisation présente quelques limites selon nous. L'amélioration principale que nous pensons possible provient du lissage du spectre que nous effectuons (4.2.3). Ce lissage permet bien de supprimer des informations non-pertinentes et d'agréger les informations prosodiques appartenant à un même niveau. En revanche, ce lissage est réalisé de manière linéaire sur les fréquences. Ainsi, deux pics sur le spectre brut situés à 0,2 Hz d'écart seraient probablement fusionnés en un seul pic. Or si ces pics se situent respectivement à 5,1 et 5,3 Hz, ils correspondent à des unités de durées de 196 ms et 185 ms, et bien la même unité prosodique. En revanche, si ces pics sont à 0,2 Hz et 0,4 Hz, ils correspondent alors à des durées de 5 secondes et 2,5 secondes ce qui peut correspondre à deux unités différentes. Ce phénomène n'est pas systématiquement problématique en fonction de l'utilisation que nous avons de l'EMS. Par exemple, celà n'est pas génant lorsque nous nous intéressons à la comparaison des régularités syllabiques par rapport aux régularités de plus haut niveaux, comme nous l'avons fait dans le chapitre 5. En revanche, pour un objectif de distinction plus précis des régularités des syntagmes intermédiaires et intonatifs, il est possible que le lissage masque de l'information. Afin d'améliorer notre méthodologie, il aurait été possible de réaliser un lissage de façon logarithmique. C'est à dire en laissant les valeurs brutes de l'EMS pour les fréquences les plus basses, et d'agréger un plus grand nombre de pics pour des fréquences plus hautes. Cela serait donc possible en modifiant notre fenêtre de lissage de 0,3 Hz (voir 4.2.3 pour plus de détails) en l'adaptant en fonction des fréquences traitées.

Une autre piste pourrait provenir peut être dans l'amélioration du spectre de modulations de fréquence, dans le but d'améliorer la précision du modèle sur les unités

prosodiques les plus larges. En effet, nous avons pu voir qu'en utilisant une modélisation de la courbe intonative comme Momel Hirst et Espesser (1993) à la place de l'enveloppe d'amplitude, il était possible de rendre compte efficacement de régularités des syntagmes intonatifs et/ou intermédiaires. Utiliser une combinaison des modulations d'amplitude et de fréquences pourrait donc être une alternative intéressante afin d'améliorer notre modélisation. Une piste viable pourrait être par exemple d'utiliser "l'énergie périodique" décrite par Albert et collab. (2018). Cette énergie périodique se base sur une combinaison de l'énergie du signal et de la f_0. Ceci pourrait donc être une bonne alternative à nos enveloppes d'amplitude et de f_0.

Améliorer la représentation en spectrogramme du rythme

Parmi les modélisations que nous avons peu exposées dans cette thèse, nous pouvons citer le spectrogramme du rythme (détaillé dans la section 4.3) qui à première vue nous semblait prometteur mais pour lequel nous n'avons pas réussi à extraire des informations fiables et reproductibles sur la parole continue. Dans l'idéal, le spectrogramme du rythme aurait pu nous permettre d'observer l'évolution des régularités rythmiques sur la parole continue à la manière du modèle théorique des oscillateurs couplés dont nous avons discuté dans la section 1.2.2. Le spectrogramme du rythme n'a cependant pas réussi à capturer ces évolutions sur des successions de phrases (tant sur du slam que de la lecture de texte). Parmi les pistes d'améliorations de modèle, nous pensons que le choix de la fenêtre d'analyse impacte fortement les résultats produits. Nous avons en effet choisi d'utiliser une fenêtre glissante de trois secondes dans nos expériences. Cependant, comme nous l'avons vu dans la section 5.2.2, le débit de parole d'une personne à une autre est très variable. Ainsi, en choisissant une taille de fenêtre fixe, le spectrogramme ne capture pas les mêmes informations d'un locuteur à l'autre. C'est pourquoi nous pensons qu'utiliser une fenêtre glissante dont la taille dépend du débit de parole serait une piste intéressante à explorer.

Le rôle des pauses dans l'équilibrage rythmique

L'un des aspects que nous n'avons pas traité et qui pourtant est un élément crucial de l'étude du rythme de la parole est le rôle des pauses. En effet, les pauses participent à priori fortement dans le cadre de l'équilibrage rythmique des unités prosodiques. C'est ainsi que les durées des pauses sont généralement dans un rapport entier par rapport à la durée du segment de parole qui les précède. Nous ne nous sommes alors pas posé la question de savoir si l'EMS permettait de rendre compte de cet équilibrage rythmique. Afin d'examiner cela, nous avons réalisé une expérience portant sur la comparaison de l'EMS d'une lecture de texte classique avec l'EMS de cette même lecture à laquelle nous avons enlevé les pauses dans le signal. Cette expérimentation est visible sur la figure 7.1. Nous pouvons alors observer sur cet exemple que la régularité syllabique est maintenue, mais la régularité des syntagmes intonatifs et/ou intermédiaires est totalement écrasée une fois les pauses retirées. Ceci suggère donc

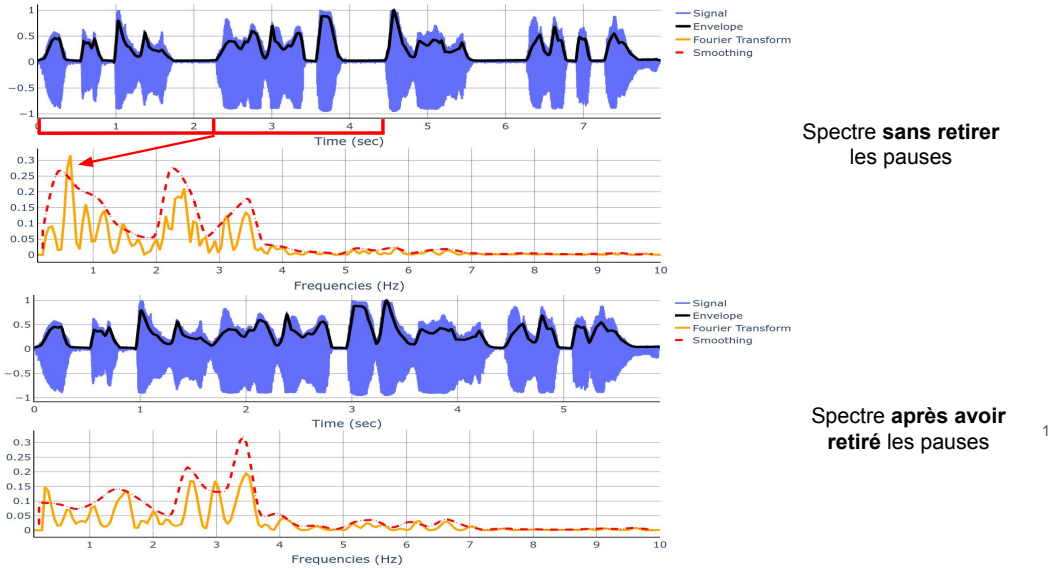

FIGURE 7.1 – Comparaison de deux spectres de modulations d'amplitude d'une personne atteinte de cancer VADS (patient n^o 308 ; sévérité = 1,3). Le premier (en haut) correspond à la lecture normale de la première phrase de la chèvre de Monsieur Seguin. Le second (en bas) correspond au même signal de lecture auquel nous avons supprimé les pauses via un détecteur d'activité vocale.

que les pauses participent en effet à l'équilibrage rythmique de ces syntagmes. De plus, cela montre que l'EMS prend bien en compte les pauses dans les mesures de régularités des unités rythmiques. Cependant, nous n'avons pas encore étendu cette expérience à l'ensemble de nos corpus. Il serait donc intéressant de généraliser ces résultats au travers de mesures d'énergies dans des bandes de basses fréquences avant et après la suppression des pauses. Nous pourrions ainsi quantifier l'équilibrage induit par les pauses et ainsi voir si leurs durées chez certains patients sont plus variables ou bien si elles permettent systématiquement de contrebalancer les durées de certains syntagmes.

Caractérisation automatique de la prosodie

Dans la section 5.3.2, nous avons proposé une méthodologie ayant pour but de caractériser automatiquement les stratégies rythmiques d'un ensemble de locuteurs. Pour cela, nous nous somme basés uniquement sur des paramètres du rythme extraits du spectre de modulations d'amplitude. Bien que notre projection soit en partie pertinente par rapport à nos annotations prosodiques, les caractéristiques de plusieurs unités rythmiques ne sont pas modélisées. Ainsi, nous nous focalisons sur des pics de l'EMS sans connaître au préalable les niveaux auxquels ils correspondent. Nous pouvons néanmoins estimer ces unités au travers des fréquences de pics relativement au débit de parole, mais il est difficile d'en être sûr. Une étude plus large des EMS

de patients en relation avec les annotations des durées des unités prosodiques serait alors envisageable. Nous pourrions par exemple déterminer les fréquences typiques des niveaux prosodiques en fonction du débit de parole et du débit articulatoire. Cela nous permettrait de voir si ces niveaux sont systématiquement dépendant du nombre de syllabes ou bien si leurs durées sont globalement les mêmes.

Dans le cadre de la caractérisation automatique de la prosodie, nous pourrions imaginer améliorer notre représentation en deux dimensions. En plus de l'extraction de nouveaux paramètres basés sur l'EMS, nous pourrions ainsi intégrer un ensemble de paramètres rythmiques plus larges. Par exemple en incluant des informations sur les valeurs et les variations de f_0, ou encore des mesures sur les durées des pauses ou du débit de parole. Il serait alors possible d'obtenir une catégorisation plus globale de la prosodie. Avec l'ajout de caractéristiques de ce type, il serait intéressant d'utiliser un algorithme de réduction de dimensions différents tel que l'Analyse en Composante Principale (ACP). En effet, en utilisant ce type de méthode, nous pourrions obtenir des regroupements de locuteurs pertinents d'un point de vue de leurs caractéristiques prosodiques via cette méthode. Il serait enfin possible d'analyser la contribution des paramètres prosodiques dans chaque dimension obtenue, ce qui permettrait de repérer facilement les spécificités des locuteurs. Il resterait à estimer la part de la prosodie dans la perception de la compréhensibilité des énoncés. En effet, les patients atteints de cancer VADS ayant subi des traitements très handicapants, peuvent se reposer sur leur maîtrise de la prosodie pour arriver à rendre plus intelligible des énoncés très perturbés au niveau acoustique/phonémique.

A

Performances brutes des algorithmes de *f0* sur la parole pathologique

TABLE A.1 – Résultats bruts des 14 algorithmes de f_0 évalués dans la section 4.4.1. VDE correspond aux erreurs de détection de voisement, GPE représente les erreurs dans valeurs de f_0 estimées, FNR et FPR sont respectivement les faux négatifs et faux positifs dans la détection de voisement. x2 et ÷2 sont respectivement les erreurs où les estimations de f_0 sont 20% plus hautes ou plus basses par rapport à nos annotations. T correspond aux sujets témoins, C aux sujets cancers VADS et P aux sujets Parkinsoniens

	VDE (%)			**GPE (%)**			**FNR (%)**			**FPR (%)**			**x2 (%)**			**÷2 (%)**		
Algorithm	**T**	**C**	**P**	**T**	**C**	**P**	**T**	**H**	**P**	**T**	**C**	**P**	**T**	**H**	**P**	**T**	**C**	**P**
ACF	3.8	4.0	3.6	0.9	10.0	2.1	**1.3**	2.4	2.4	2.5	1.7	1.1	0.1	**0.7**	0.2	0.8	9.3	1.9
AMDF	4.5	4.7	4.6	4.6	7.2	2.6	3.1	3.6	3.6	1.4	1.0	1.0	2.1	4.6	1.4	2.4	2.7	1.2
REAPER	3.6	**3.5**	**3.3**	5.8	13.7	6.3	1.5	**1.3**	**1.7**	2.0	2.2	1.6	0.2	1.4	0.2	5.6	12.3	6.2
RAPT	5.6	4.9	5.9	1.2	8.6	2.6	2.6	2.2	**1.7**	3.0	2.7	4.2	0.2	1.4	0.5	1.0	7.3	2.1
Enhanced RAPT	12.4	9.6	6.0	**0.2**	5.3	1.3	8.1	6.7	3.3	4.2	2.9	2.7	0.1	**0.9**	0.6	**0.1**	4.4	0.7
Yin	11.7	10.1	9.1	2.3	9.8	2.8	6.8	6.6	6.1	4.9	3.5	3.0	0.1	1.2	0.2	2.2	8.5	2.6
NDF	8.0	8.4	7.3	0.6	**3.0**	0.9	1.4	3.2	2.5	6.7	5.2	4.8	0.2	1.9	0.5	0.4	**1.0**	**0.3**
YAAPT	5.0	5.9	4.6	1.3	5.0	4.0	2.7	4.4	2.9	2.3	1.4	1.6	0.2	1.7	0.3	1.1	3.2	3.8
SWIPE	6.7	9.2	6.2	79.2	68.1	91.8	6.0	8.8	5.1	**0.7**	**0.4**	1.1	18.3	5.4	9.1	60.9	62.7	82.7
PEFAC	8.1	6.6	9.5	16.8	16.7	10.5	4.1	3.5	3.9	4.0	3.0	5.5	6.9	8.0	7.1	9.9	8.6	3.4
CREPE	8.2	7.8	10.6	0.8	4.8	1.6	2.9	4.4	4.7	5.3	3.4	5.9	0.4	2.1	0.7	0.4	2.7	0.9
FCN-F0	6.5	7.7	9.6	0.3	3.8	**0.5**	3.7	5.0	5.2	2.8	2.6	4.3	**0.0**	1.7	**0.1**	0.3	2.1	0.4
Median	**3.1**	3.9	**3.3**	0.9	5.7	1.7	1.9	2.8	2.1	1.2	1.1	1.2	**0.0**	1.5	0.3	0.9	4.2	1.4
Combi	3.6	**3.5**	**3.3**	0.7	5	1.6	2.6	4.0	3.6	1.4	1.2	**0.9**	0.1	1.9	0.5	0.3	**1.0**	**0.3**

B

Exemples d'EMS sur une sélection de patients et témoins

Cette annexe contient des comparaisons entre nos trois groupes de locuteurs (contrôles, cancer VADS, MDP) sur les deux premières phrases du texte de la chèvre de Monsieur Seguin.

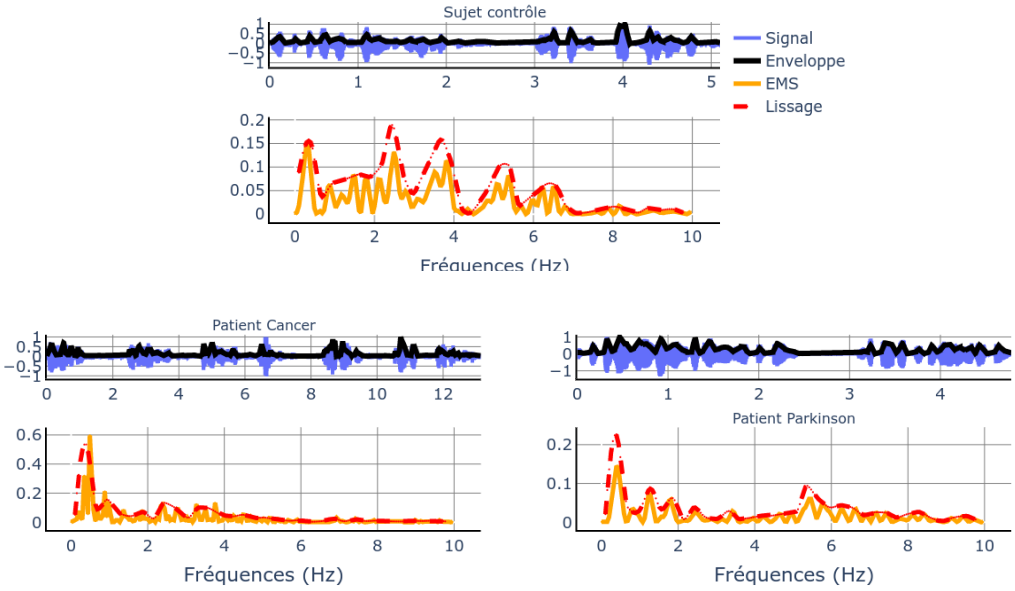

FIGURE B.1 – Le sujet contrôle (locuteur n^o 032 ; sévérité = 9,5) est en haut, l'EMS du sujet cancer VADS (locuteur n^o 338 ; sévérité = 4,2) est en bas à gauche, celui du patient Parkinson (locuteur n^o 1015 ; sévérité = 7) est en bas à droite.

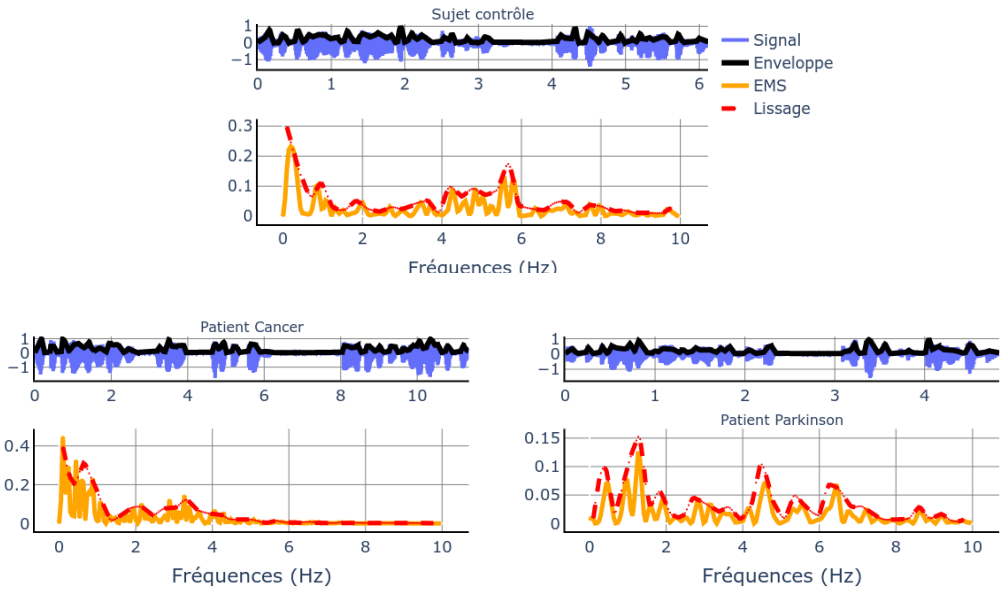

FIGURE B.2 – Le sujet contrôle (locuteur n^o 001 ; sévérité = 10) est en haut, l'EMS du sujet cancer VADS (locuteur n^o 321 ; sévérité = 3,5) est en bas à gauche, celui du patient Parkinson (locuteur n^o 1019 ; sévérité = 6,5) est en bas à droite.

C

Exemples d'EMS superposés aux unités prosodiques

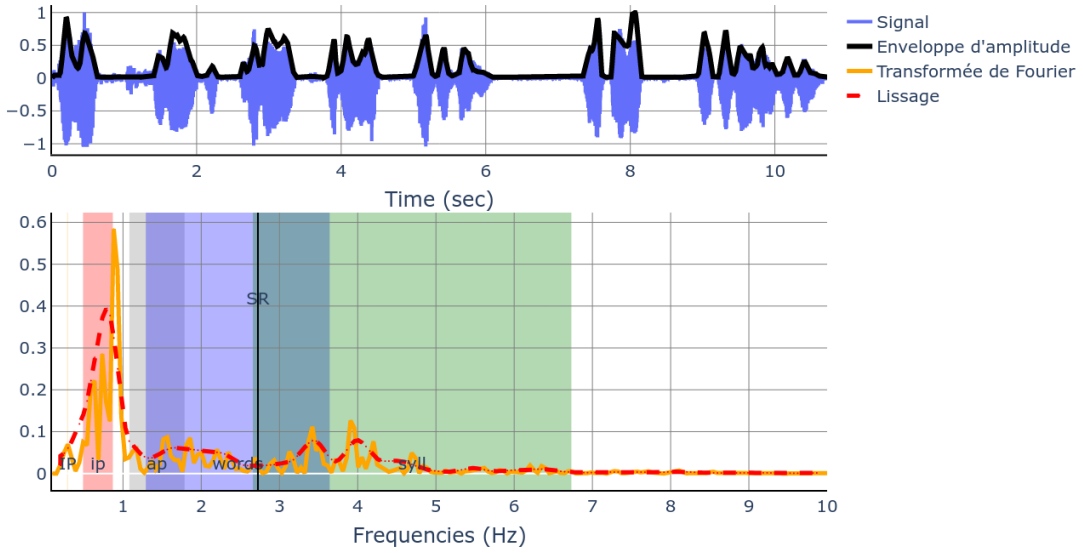

FIGURE C.1 – EMS d'un locuteur cancer VADS (locuteur n^o 301 ; sévérité = 2,6) superposé aux annotations prosodiques sur les phrases "*Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon*".

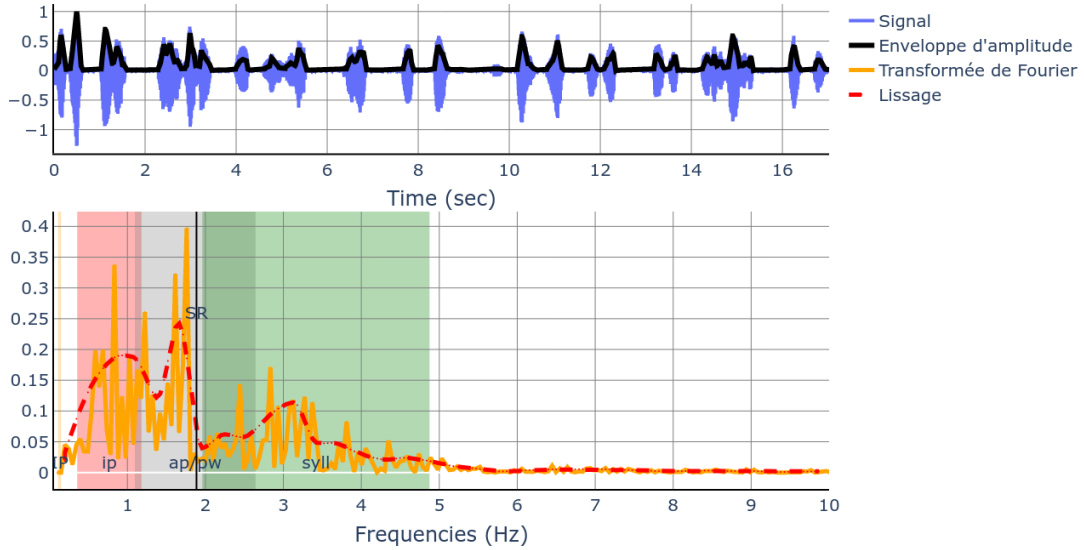

FIGURE C.2 – EMS d'un locuteur cancer VADS (locuteur n^o 304 ; sévérité = 1,8) superposé aux annotations prosodiques sur les phrases "*Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon*".

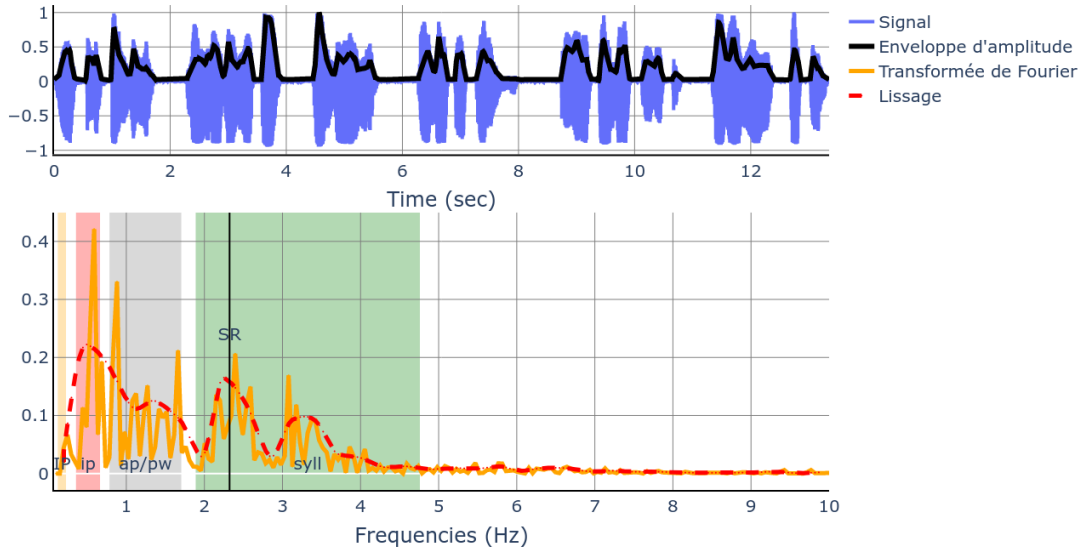

FIGURE C.3 – EMS d'un locuteur cancer VADS (locuteur n^o 308 ; sévérité = 1,3) superposé aux annotations prosodiques sur les phrases "*Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon*".

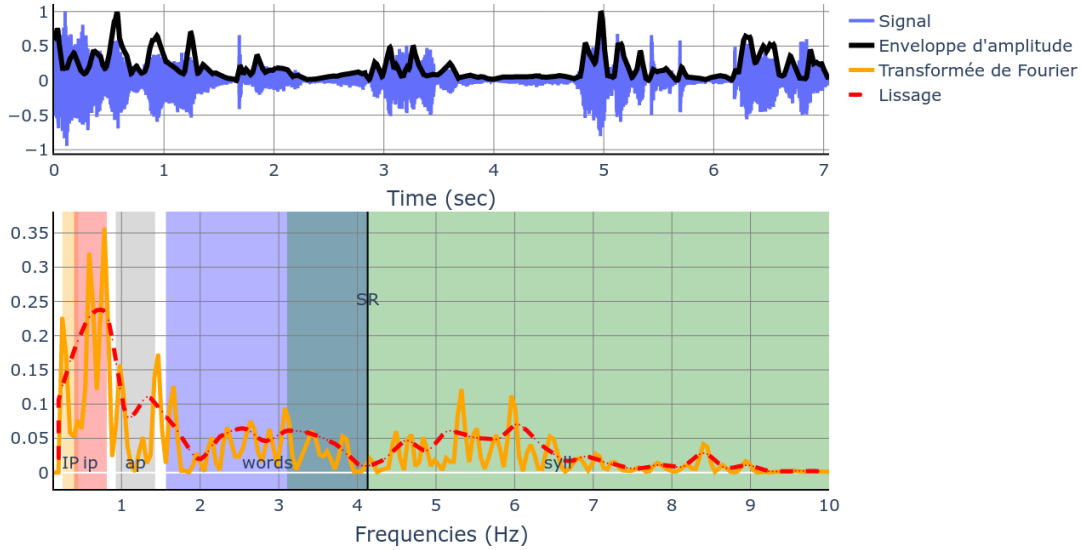

FIGURE C.4 – EMS d'un locuteur cancer VADS (locuteur n^o 330 ; sévérité = 5,4) superposé aux annotations prosodiques sur les phrases "*Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon*".

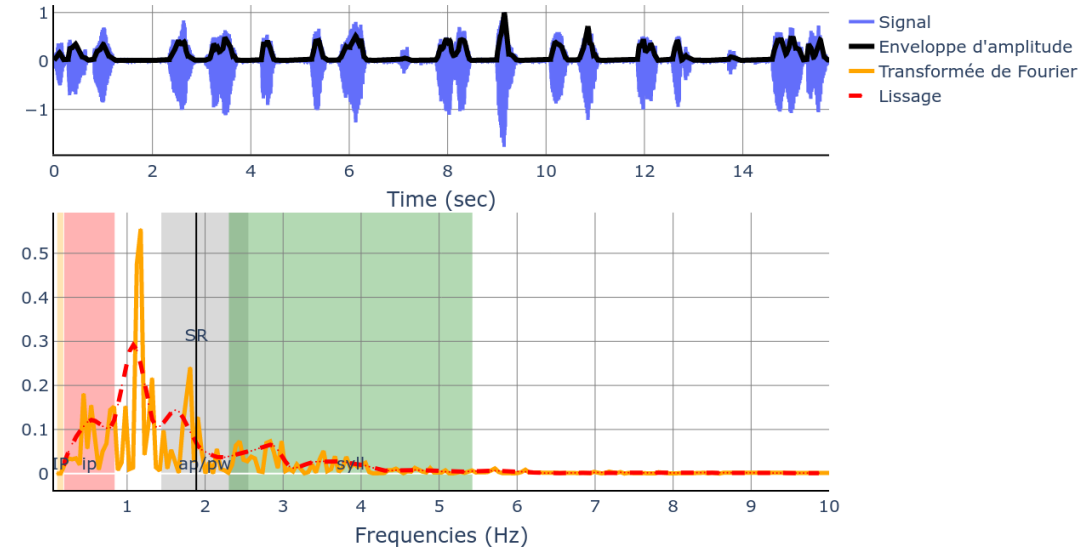

FIGURE C.5 – EMS d'un locuteur cancer VADS (locuteur n^o 353 ; sévérité = 1,2) superposé aux annotations prosodiques sur les phrases "*Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon*".

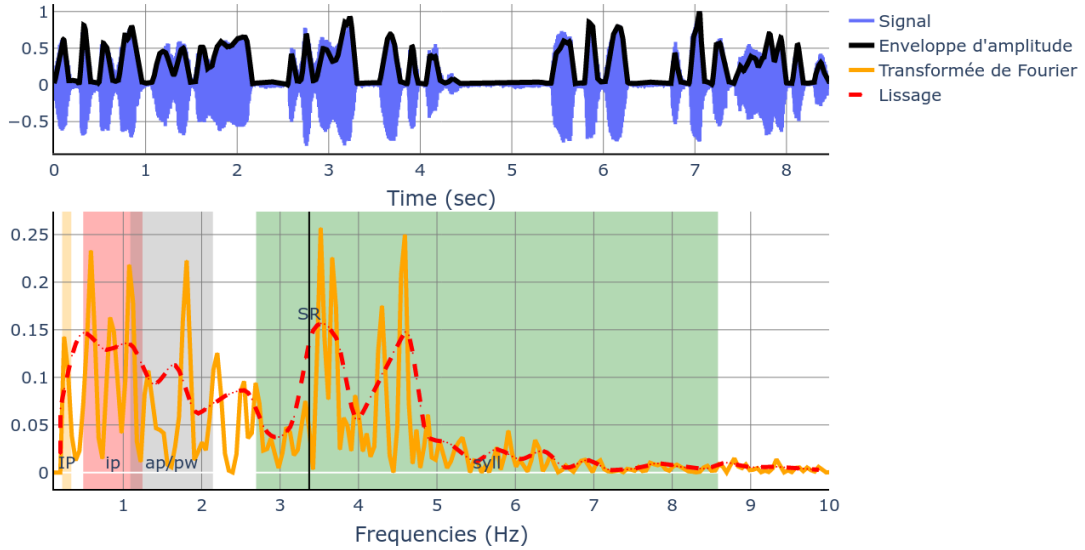

FIGURE C.6 – EMS d'un locuteur cancer VADS (locuteur n^o 355 ; sévérité = 2,7) superposé aux annotations prosodiques sur les phrases "*Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon*".

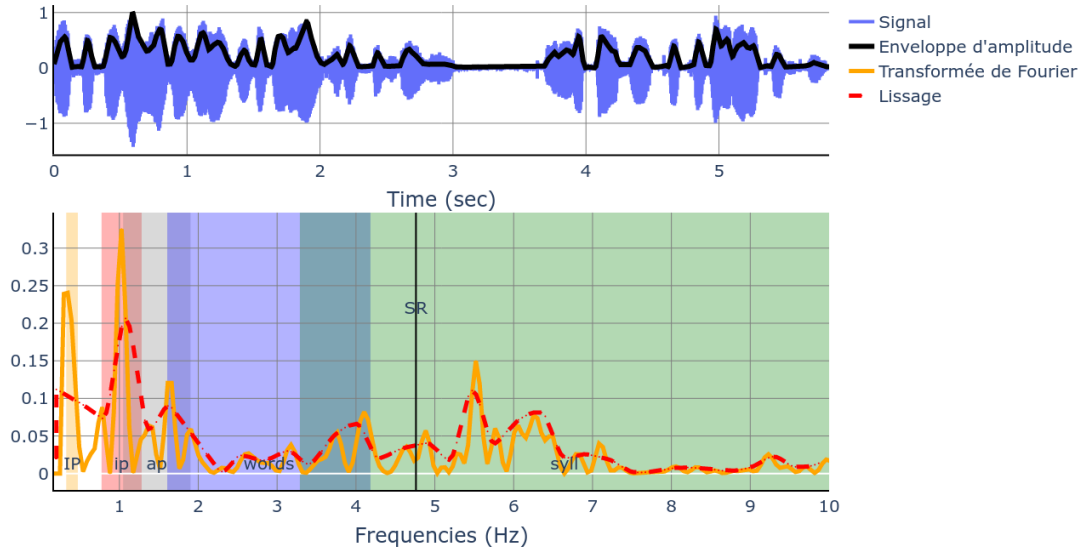

FIGURE C.7 – EMS d'un locuteur cancer VADS (locuteur n^o 362 ; sévérité = 8,8) superposé aux annotations prosodiques sur les phrases "*Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon*".

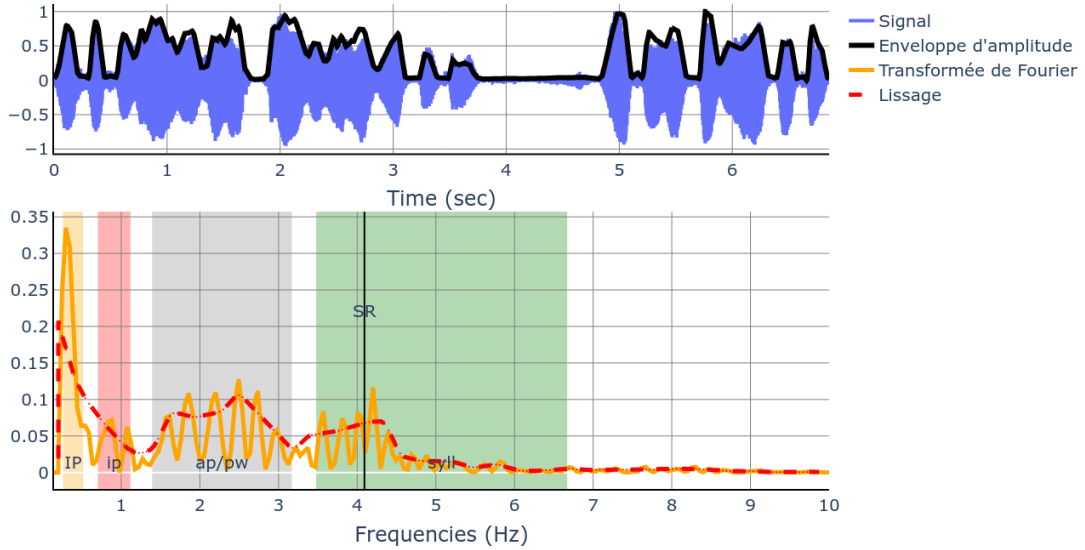

FIGURE C.8 – EMS d'un locuteur cancer VADS (locuteur n^o 363 ; sévérité = 3,8) superposé aux annotations prosodiques sur les phrases "*Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon*".

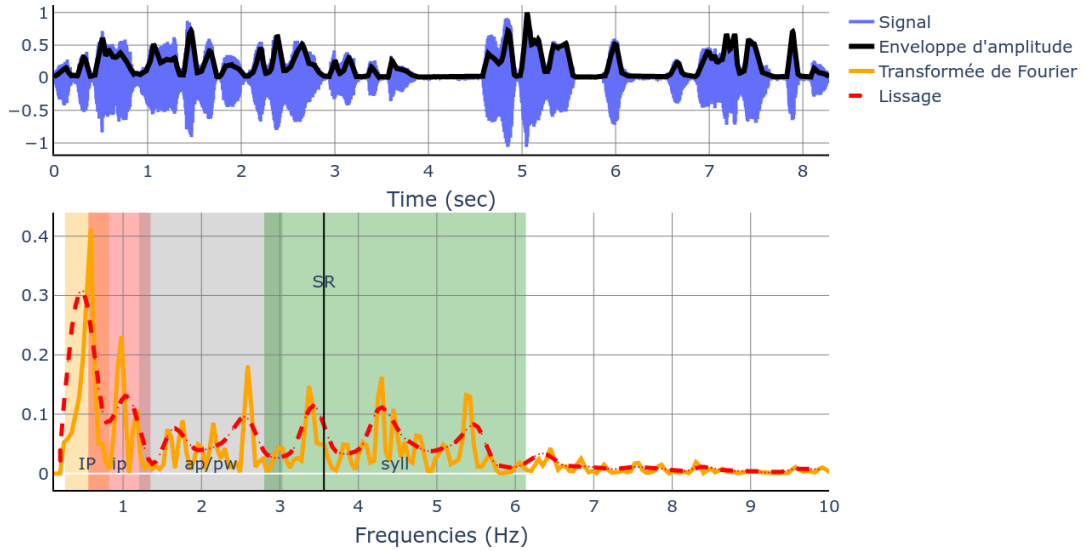

FIGURE C.9 – EMS d'un locuteur cancer VADS (locuteur n^o 392 ; sévérité = 1,7) superposé aux annotations prosodiques sur les phrases "*Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres. Il les perdait toutes de la même façon*".

Bibliographie

Akima, H. 1970, «A new method of interpolation and smooth curve fitting based on local procedures», *Journal of the ACM (JACM)*, vol. 17, n° 4, p. 589–602.

Albert, A., F. Cangemi et M. Grice. 2018, «Using periodic energy to enrich acoustic representations of pitch in speech: A demonstration», dans *Proc. Speech Prosody 2018*, p. 804–808, doi :10.21437/SpeechProsody.2018-162.

Ambikairajah, E., H. Li, L. Wang, B. Yin et V. Sethu. 2011, «Language identification: A tutorial», *IEEE Circuits and Systems Magazine*, vol. 11, n° 2, p. 82–108.

Andre-Obrecht, R. 1988, «A new statistical approach for the automatic segmentation of continuous speech signals», *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, n° 1, p. 29–40.

Arai, T. et S. Greenberg. 1997, «The temporal properties of spoken Japanese are similar to those of English», dans *Proc. 5th European Conference on Speech Communication and Technology (Eurospeech 1997)*, p. 1011–1014, doi :10.21437/Eurospeech.1997-355.

Ardaillon, L. et A. Roebel. 2019, «Fully-Convolutional Network for Pitch Estimation of Speech Signals», dans *Proc. Interspeech 2019*, p. 2005–2009, doi :10.21437/Interspeech.2019-2815.

Arvaniti, A. 2021, «Measuring speech rhythm», dans *The Cambridge Handbook of Phonetics*, édité par R.-A. Knight et J. Setter, Cambridge Handbooks in Language and Linguistics, Cambridge University Press, p. 312–335, doi :10.1017/9781108644198.013.

Astésano, C. 2001, *Rythme et accentuation en français : invariance et variabilité stylistique*, Editions L'Harmattan.

Astésano, C. 2016, «The Prosodic Characterization of Reference French», dans *Varieties of Spoken French*, édité par S. Detey, J. Durand, B. Laks et C. Lyche, Oxford University Press, p. 68–85. URL `https://hal-univ-tlse2.archives-ouvertes.fr/hal-02159859`.

Astésano, C. 2017, «Le statut de l'accent initial dans la phonologie prosodique du français : enjeux descriptifs et psycholinguistiques», *Habilitation à diriger des Recherches, UT2J*.

Astésano, C. et R. Bertrand. 2016, «Accentuation et niveaux de constituance en français : enjeux phonologiques et psycholinguistiques», *Langue française*, vol. 191, n° 3, p. 11–30.

Astésano, C., R. Bertrand, R. Espesser et N. Nguyen. 2012, «Perception des frontières et des proéminences en français», dans *Journées d'études sur la Parole*, Grenoble, France, p. 8. URL https://hal.archives-ouvertes.fr/hal-01510445.

Astésano, C., C. Magne, R. El Yagoubi et M. Besson. 2004, «Influence du rythme sur le traitement sémantique en français : approches comportementale et électrophysiologique», dans *Actes des XXVe Journées d'Études de la Parole (JEP 2004), Fès*.

Astésano, C. 1998, «Effects of prosodic constraints on the differential lengthening of syllable constituents in french: A comparison between spontaneous and read speech», dans *ISCA Workshop on Sound Patterns of Spontaneous Speech*, p. 143–146.

Astésano, C. 2019, «The prosodic word as the domain of french accentuation - empirical evidence», dans *Phonetics and Phonology in Europe (PaPE) 2019*, p. 170–171.

Astésano, C. 2022, «De la supramodalité du rythme : Implications pour la description prosodique, la remédiation linguistique et l'apprentissage des langues», dans *Proc. XXXIVe Journées d'Études sur la Parole - JEP 2022*, p. 1–14, doi :10.21437/JEP.2022-1.

Astésano, C., A. Di Cristo et D. Hirst. 1995, «Discourse-based empirical evidence for a multi-class accent system in french», *XIIIème Congrès International des Sciences Phonétiques*, vol. 4.

Aura, K. 2012, *Protocole d'évaluation du langage fondé sur le traitement de fonctions prosodiques : étude exploratoire de deux patients atteints de gliomes de bas grade en contexte péri-opératoire*, thèse de doctorat, Université Toulouse le Mirail-Toulouse II.

Balaguer, M., A. Boisguérin, A. Galtier, N. Gaillard, M. Puech et V. Woisard. 2019a, «Assessment of impairment of intelligibility and of speech signal after oral cavity and oropharynx cancer», *European Annals of Otorhinolaryngology, Head and Neck Diseases*, vol. 136, n° 5, doi :https://doi.org/10.1016/j.anorl.2019.05.012, p. 355–359, ISSN 1879-7296. URL https://www.sciencedirect.com/science/article/pii/S1879729619301024.

Balaguer, M., A. Boisguérin, A. Galtier, N. Gaillard, M. Puech et V. Woisard. 2019b, «Jugement d'altération de l'intelligibilité et de sévérité d'un trouble de la production de la parole séquellaire d'un cancer de la cavité buccale ou de l'oropharynx», *Annales françaises d'Oto-rhino-laryngologie et de Pathologie Cervico-faciale*, vol. 136, n° 5, doi :https://doi.org/10.1016/j.aforl.2019.01.002, p. 347–352, ISSN 1879-7261. URL https://www.sciencedirect.com/science/article/pii/S1879726119301160.

Bayestehtashk, A., M. Asgari, I. Shafran et J. McNames. 2015, «Fully automated assessment of the severity of parkinson's disease from speech», *Computer speech & language*, vol. 29, n° 1, p. 172–185.

Beckman, M. 1996, «The parsing of prosody», *Language and Cognitive Processes - LANG COGNITIVE PROCESS*, vol. 11, doi :10.1080/016909696387213, p. 17–68.

Beckman, M. E. 1986, «Stress and non-stress accent», *Netherlands phonetic archives*.

Beckman, M. E. 1992, «Evidence for speech rhythms across languages», *Speech perception, production and linguistic structure*, p. 457–463.

Beckman, M. E. et J. B. Pierrehumbert. 1986, «Intonational structure in japanese and english», *Phonology*, vol. 3, doi :10.1017/S095267570000066X, p. 255–309.

Benguerel, A.-P. 1973, «Corrélats physiologiques de l'accent en français», *Phonetica*, vol. 27, nº 1, p. 21–35.

Bertrand, R., C. Astésano et N. Nguyen. 2019, «Prominence and boundary are two distinct phenomena in french: perceptual evidence», dans *Phonetics and Phonology in Europe (PaPE) 2019*, p. 51–52.

Boersma, P. 2000, «Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound», *Proceedings of the Institute of Phonetic Sciences*, vol. 17.

Boersma, P. 2001, «Praat, a system for doing phonetics by computer», *Glot International*, vol. 5, nº 9, p. 341–345.

Bracewell, R. N. et R. N. Bracewell. 1986, *The Fourier transform and its applications*, vol. 31999, McGraw-Hill New York.

Brown, J. C. 1993, «Determination of the meter of musical scores by autocorrelation», *The Journal of the Acoustical Society of America*, vol. 94, nº 4, p. 1953–1957.

Butterworth, S. et collab.. 1930, «On the theory of filter amplifiers», *Wireless Engineer*, vol. 7, nº 6, p. 536–541.

Cai, W., D. Cai, S. Huang et M. Li. 2019, «Utterance-level end-to-end language identification using attention-based cnn-blstm», dans *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 5991–5995.

Cai, W., Z. Cai, X. Zhang, X. Wang et M. Li. 2018, «A novel learnable dictionary encoding layer for end-to-end language identification», dans *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, p. 5189–5193.

Camacho, A. et J. Harris. 2008, «A sawtooth waveform inspired pitch estimator for speech and music», *The Journal of the Acoustical Society of America*, vol. 124, doi : 10.1121/1.2951592, p. 1638–1652.

de Cheveigne, A. et H. Kawahara. 2002, «YIN, a fundamental frequency estimator for speech and musica)», *Journal of the Acoustical Society of America*, vol. 111, nº 4, p. 14.

Chi, T., Y. Gao, M. C. Guyton, P. Ru et S. Shamma. 1999, «Spectro-temporal modulation transfer functions and speech intelligibility», *The Journal of the Acoustical Society of America*, vol. 106, nº 5, doi :10.1121/1.428100, p. 2719–2732. URL `https://doi.org/10.1121/1.428100`.

Couper-Kuhlen, E. 1993, «English speech rhythm», *English Speech Rhythm*, p. 1–360.

Cummins, F. et R. Port. 1998a, «Rhythmic constraints on stress timing in english», *Journal of Phonetics*, vol. 26, nº 2, p. 145–171.

Cummins, F. et R. Port. 1998b, «Rhythmic constraints on stress timing in English», *Journal of Phonetics*, vol. 26, n° 2, doi :10.1006/jpho.1998.0070, p. 145–171, ISSN 00954470. URL https://linkinghub.elsevier.com/retrieve/pii/S0095447098900705.

Cutler, A., D. Dahan et W. Van Donselaar. 1997, «Prosody in the comprehension of spoken language: A literature review», *Language and speech*, vol. 40, doi :10.1177/002383099704000203, p. 141–201.

Darley, F. L., A. E. Aronson et J. R. Brown. 1969a, «Clusters of deviant speech dimensions in the dysarthrias», *Journal of Speech and Hearing Research*, vol. 12, n° 3, doi : 10.1044/jshr.1203.462, p. 462–496. URL https://pubs.asha.org/doi/abs/10.1044/jshr.1203.462.

Darley, F. L., A. E. Aronson et J. R. Brown. 1969b, «Differential diagnostic patterns of dysarthria», *Journal of Speech and Hearing Research*, vol. 12, n° 2, doi :10.1044/jshr.1202.246, p. 246–269. URL https://pubs.asha.org/doi/abs/10.1044/jshr.1202.246.

Daubechies, I. 1988, «Orthonormal bases of compactly supported wavelets», *Communications on pure and applied mathematics*, vol. 41, n° 7, p. 909–996.

Daudet, A. 1870, *Lettres de mon moulin : impressions et souvenirs*, Hetzel.

Dauer, R. M. 1983, «Stress-timing and syllable-timing reanalyzed», *Journal of phonetics*, vol. 11, n° 1, p. 51–62.

Davis, S. et P. Mermelstein. 1980, «Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences», *IEEE transactions on acoustics, speech, and signal processing*, vol. 28, n° 4, p. 357–366.

De Cornulier, B. 1995, «Métrique», *Encyclopédia Universalis. Paris : Hachette*.

De Jong, N. H. et T. Wempe. 2009, «Praat script to detect syllable nuclei and measure speech rate automatically», *Behavior research methods*, vol. 41, n° 2, p. 385–390.

Delais-Roussarie, E. 1995, *Pour une approche parallèle de la structure prosodique : étude de l'organisation prosodique et rythmique de la phrase française*, thèse de doctorat, Toulouse 2.

Delattre, P. 1966, «Les dix intonations de base du français», *French review*, p. 1–14.

Dellwo, V. 2006, «Rhythm and speech rate: A variation coefficient for deltac», dans *Language and language-processing*, édité par P. Karnowski et I. Szigeti, Peter Lang, Frankfurt, ISBN 978-3-631-55477-7, p. 231–241. URL https://doi.org/10.5167/uzh-111789.

Dellwo, V. et P. Wagner. 2015, «Relations between language rhythm and speech rate», dans *Proceedings of the International Congress of Phonetics Science. International Congress of Phonetics Science*, p. 471–474.

Desain, P. 1992, «A (de) composable theory of rhythm perception», *Music perception*, vol. 9, n° 4, p. 439–454.

Di Cristo, A. 1999, «Le cadre accentuel du français contemporain : essai de modélisation. première partie», *Langues (Montrouge)*, ISSN 1291-1542.

Di Cristo, A. 2000, «Vers une modélisation de l'accentuation du français (seconde partie)», *Journal of French language studies*, vol. 10, n° 1, p. 27–44.

Di Cristo, A. 2004, «La prosodie au carrefour de la phonétique, de la phonologie et de l'articulation formes-fonctions», *Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix-en-Provence (TIPA)*, vol. 23, p. 67–211.

Di Cristo, A. 2011, «Une approche intégrative des relations de l'accentuation au phrasé prosodique du français», *Journal of French Language Studies*, vol. 21, n° 1, doi :10.1017/S0959269510000505, p. 73—95.

Di Cristo, A. 2016a, *Les Musiques du Français Parlé*, De Gruyter, 440 p..

Di Cristo, A. 2016b, *Les musiques du français parlé : Essais sur l'accentuation, la métrique, le rythme, le phrasé prosodique et l'intonation du français contemporain*, de Gruyter.

Di Cristo, A. et D. Hirst. 1996, «Vers une typologie des unités intonatives du français», *XXIème JEP, Avignon*, p. 219–22.

D'Imperio, M. et A. Michelas. 2010, «Embedded register levels and prosodic phrasing in french», dans *Speech Prosody*, p. 4.

Dirac, P. A. M. 1930, *The principles of quantum mechanics*, vol. 27, The Clarendon Press, Oxford.

Dittner, J., B. Lepage, V. Woisard, M. Kergadallan, K. Boisteux, E. Robart et M. Welby-Gieusse. 2010, «Elaboration et validation d'un test quantitatif d'intelligibilitépour les troubles pathologiques de la production de la parole», *Rev Laryngol Otol Rhinol (Bord)*, vol. 131, n° 1, p. 9–14, ISSN 0035-1334 (Print) ; 0035-1334 (Linking).

Dixon, S. 2001, «Automatic extraction of tempo and beat from expressive performances», *Journal of New Music Research*, vol. 30, n° 1, p. 39–58.

Drucker, H., C. J. C. Burges, L. Kaufman, A. Smola et V. Vapnik. 1997, «Support vector regression machines», dans *Advances in Neural Information Processing Systems*, vol. 9, édité par M. C. Mozer, M. Jordan et T. Petsche, MIT Press, p. 155–161.

Dupoux, E., C. Pallier, N. Sebastian et J. Mehler. 1997, «A destressing "deafness" in french?», *Journal of Memory and Language*, vol. 36, n° 3, p. 406–421.

D'Imperio, M. 2002, «Italian intonation: An overview and some questions», *Probus*, vol. 14, p. 37–69.

Elliott, T. M. et F. E. Theunissen. 2009, «The modulation transfer function for speech intelligibility», *PLOS Computational Biology*, vol. 5, n° 3, doi :10.1371/journal.pcbi.1000302, p. 1–14. URL `https://doi.org/10.1371/journal.pcbi.1000302`.

Fant, G., A. Kruckenberg et L. Nord. 1991, «Durational correlates of stress in swedish, french and english», *Journal of Phonetics*, vol. 19, p. 351–365.

Farinas, J. 2002, *Une modélisation automatique du rythme pour l'identification des langues* , Thèse de doctorat, Université Paul Sabatier, Toulouse, France. (Soutenance le 15/11/2002).

Farinas, J. et F. Pellegrino. 2001, «Automatic rhythm modeling for language identification», dans *Seventh European Conference on Speech Communication and Technology*, p. 2539–2542.

Farinas, J., J.-L. Rouas, F. Pellegrino et R. André-Obrecht. 2005, «Extraction automatique de paramètres prosodiques pour l'identification automatique des langues», *Traitement du Signal*, vol. 22, n° 2, p. 81–97. ISSN : 0765-0019.

Fletcher, J. 1991, «Rhythm and final lengthening in french», *Journal of phonetics*, vol. 19, n° 2, p. 193–212.

Fónagy, I. 1980, «L'accent français : accent probabilitaire (dynamique d'un changement prosodique)», *Studia Phonetica Montréal*, vol. 15, p. 123–233.

Fougeron, C. et S.-A. Jun. 1998, «Rate effects on french intonation: prosodic organization and phonetic realization», *Journal of Phonetics*, vol. 26, n° 1, doi : https://doi-org-s.docadis.univ-tlse3.fr/10.1006/jpho.1997.0062, p. 45–69, ISSN 0095-4470. URL https://www-sciencedirect-com-s.docadis.univ-tlse3.fr/science/article/pii/S0095447097900620.

Fourakis, M. 1991, «Tempo, stress, and vowel reduction in american english», *The Journal of the Acoustical society of America*, vol. 90, n° 4, p. 1816–1827.

Fraisse, P. 1974, *Psychologie du rythme*, Presse Universitaire de France.

Garde, P. 1968, *L'accent*, vol. 5, Presses universitaires de France.

Garnier, L. 2018, *Quels liens entre accentuation et niveaux de constituance en français ? : une analyse perceptive et acoustique*, thèse de doctorat, Université Toulouse le Mirail-Toulouse II. Non publiée.

Garnier, L., L. Baqué, A. Dagnac et C. Astésano. 2016, «Perceptual investigation of prosodic phrasing in French», dans *Speech Prosody 2016*, Speech Prosody 2016, Boston, United States, p. 1153–1157. URL https://hal.science/hal-01330866.

Ghahremani, P., B. BabaAli, D. Povey, K. Riedhammer, J. Trmal et S. Khudanpur. 2014, «A pitch extraction algorithm tuned for automatic speech recognition», dans *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, ISBN 978-1-4799-2893-4, p. 2494–2498, doi :10.1109/ICASSP.2014.6854049.

Ghio, A., G. Pouchoulin, B. Teston, S. Pinto, C. Fredouille, C. De Looze, D. Robert, F. Viallet et A. Giovanni. 2012, «How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers?», *Speech Communication*, vol. 54, n° 5, doi :10.1016/j.specom.2011.04.002, p. 664–679.

Ghosh, B. K. 2002, «Probability inequalities related to markov's theorem», *The American Statistician*, vol. 56, n° 3, p. 186–190, ISSN 00031305. URL http://www.jstor.org/stable/3087296.

Gibbon, D. 2021, «The rhythms of rhythm», *Journal of the International Phonetic Association*, doi :10.1017/S0025100321000086, p. 1–33.

Goberman, A. M. et C. A. Coelho. 2002, «Acoustic analysis of parkinsonian speech i: speech characteristics and l-dopa therapy.», *NeuroRehabilitation*, vol. 17, nº 3, p. 237–46.

Goldman, J.-P. 2011, «Easyalign: an automatic phonetic alignment tool under praat», dans *Interspeech 2011, Firenze, Italy*.

Goldman, J.-P., A. Auchlin et A. C. Simon. 2009, «Discrimination de styles de parole par analyse prosodique semi-automatique», dans *Interface Discours & Prosodie (IDP2019), Paris*, p. 207–221.

Goldsmith, J. 1976, *Autosegmental phonology*, thèse de doctorat, MIT Press London.

Gonzalez, S. et M. Brookes. 2014, «Pefac - a pitch estimation algorithm robust to high levels of noise», *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, nº 2, p. 518–530.

Google. 2011, «Webrtc - real-time communication for the web», URL `https://webrtc.org/`, accessed : 2023-01-10.

Google-Open-Source. 2015, «Reaper: Robust epoch and pitch estimator», URL `https://github.com/google/REAPER`, accessed : 2022-09-20.

Gouyon, F. et S. Dixon. 2005, «A review of automatic rhythm description systems», *Computer music journal*, vol. 29, nº 1, p. 34–54.

Grabe, E. et E. L. Low. 2002, «Durational variability in speech and the rhythm class hypothesis», *Papers in laboratory phonology*, vol. 7, nº 1982, p. 515–546.

Grosche, P., M. Müller et F. Kurth. 2010, «Cyclic tempogram—a mid-level tempo representation for musicsignals», dans *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, IEEE, p. 5522–5525.

Guibert, M., B. Lepage, V. Woisard, M. Rives, E. Serrano et S. Vergez. 2011, «Quality of life in patients treated for advanced hypopharyngeal or laryngeal cancer», *Eur Ann Otorhinolaryngol Head Neck Dis*, vol. 128, nº 5, doi :10.1016/j.anorl.2011.02.010, p. 218–223, ISSN 1879-730X (Electronic) ; 1879-7296 (Linking).

Halle, M. et J.-R. Vergnaud. 1987, *An essay on stress*, MIT press Cambridge, MA.

Hayes, B. 1995, *Metrical stress theory: Principles and case studies*, University of Chicago Press.

Hirst, D., C. Astésano et A. Di Cristo. 1998, «Differential lengthening of syllabic constituents in french: the effect of accent type and speaking style.», dans *ICSLP*.

Hirst, D. et R. Espesser. 1993, «Automatic modelling of fundamental frequency using a quadratic spline function.», *Travaux de l'Institut de Phonétique d'Aix*, vol. 15, p. 71–85.

Hixon, T. J., G. Weismer et J. D. Hoit. 2008, *Preclinical speech science: Anatomy, physiology, acoustics, and perception*, Plural Publishing.

Hjelmslev, L. 1937, *Accent, intonation, quantité*, Istituto per l'europa orientale.

Huang, N. E., Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung et H. H. Liu. 1998, «The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis», *Proceedings of the Royal Society of London. Series A : Mathematical, Physical and Engineering Sciences*, vol. 454, nº 1971, doi :10.1098/rspa.1998.0193, p. 903–995. URL `https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1998.0193`.

Jang, S., S. Choi, H. Kim, H. Choi et Y. Yoon. 2007, «Evaluation of performance of several established pitch detection algorithms in pathological voices», dans *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, p. 620–623, doi :10.1109/IEMBS.2007.4352366.

Jankowski, L., C. Astésano et A. Di Cristo. 1999, «The initial rhythmic accent in french: Acoustic data and perceptual investigation», dans *Proceedings of the 14th International Congress of Phonetic Sciences*, vol. 1, p. 257–260.

Jeancolas, L., D. Petrovska-Delacrétaz, S. Lehéricy, H. Benali et B. Benkelfat. 2016, «L'analyse de la voix comme outil de diagnostic précoce de la maladie de parkinson : état de l'art», dans *CORESA*, vol. 2016, p. 18e.

Jouvet, D. et Y. Laprie. 2017, «Performance analysis of several pitch detection algorithms on simulated and real noisy speech data», dans *2017 25th European Signal Processing Conference (EUSIPCO)*, p. 1614–1618, doi :10.23919/EUSIPCO.2017.8081482.

Jun, S.-A. et C. Fougeron. 2000, «A phonological model of french intonation», dans *Intonation*, Springer, p. 209–242.

Jun, S.-A. et C. Fougeron. 2002, «The realizations of the accentual phrase in french intonation», *Probus*, vol. 14, doi :10.1515/prbs.2002.002, p. 147–172.

Kasi, K. et S. Zahorian. 2002, «Yet another algorithm for pitch tracking», dans *2002 IEEE international Conference on Acoustics, Speech, and Signal Processing*, vol. 1, ISBN 0-7803-7402-9, p. 361, doi :10.1109/ICASSP.2002.5743729.

Kawahara, H., A. Cheveigné, H. Banno, T. Takahashi et T. Irino. 2005, «Nearly defect-free f0 trajectory extraction for expressive speech modifications based on straight», dans *Ninth European Conference on Speech Communication and Technology*, p. 537–540.

Kim, J. W., J. Salamon, P. Li et J. P. Bello. 2018, «CREPE: A Convolutional Representation for Pitch Estimation», *arXiv :1802.06182 [cs, eess, stat]*. ArXiv : 1802.06182.

Large, E. W. et J. F. Kolen. 1994, «Resonance and the perception of musical meter», *Connection Science*, vol. 6, nº 2-3, doi :10.1080/09540099408915723, p. 177–208.

Le Coz, M. 2014, *Spectre de rythme et sources multiples : au cœur des contenus ethnomusicologiques et sonores*, thèse de doctorat, Université Toulouse 3, Paul Sabatier.

Le Coz, M., H. Lachambre, L. Koenig et R. Andre-Obrecht. 2010, «A segmentation-based tempo induction method.», dans *The International Society for Music Information Retrieval*, p. 27–32.

Léon, P. 2011, *Phonétisme et prononciations du français*, Armand Colin.

Leong, V. et U. Goswami. 2014a, «Impaired extraction of speech rhythm from temporal modulation patterns in speech in developmental dyslexia», *Frontiers in Human Neuroscience*, vol. 8, ISSN 1662-5161. URL 10.3389/fnhum.2014.00096, publisher : Frontiers.

Leong, V. et U. Goswami. 2014b, «Impaired extraction of speech rhythm from temporal modulation patterns in speech in developmental dyslexia», *Frontiers in Human Neuroscience*, vol. 8, doi :10.3389/fnhum.2014.00096, ISSN 1662-5161. URL https://www.frontiersin.org/articles/10.3389/fnhum.2014.00096.

Liberman, M. et A. Prince. 1977, «On stress and linguistic rhythm», *Linguistic inquiry*, vol. 8, nº 2, p. 249–336.

Liberman, M. Y. 1975, *The intonational system of English.*, thèse de doctorat, Massachusetts Institute of Technology.

Ling, L. E., E. Grabe et F. Nolan. 2000, «Q uantitative characterizations of speech rhythm: Syllable-timing in singapore english», *Language and speech*, vol. 43, nº 4, p. 377–401.

Liss, J. M., S. LeGendre et A. J. Lotto. 2010, «Discriminating dysarthria type from envelope modulation spectra», *Journal of Speech, Language, and Hearing Research*, vol. 53, nº 5, doi :10.1044/1092-4388(2010/09-0121), p. 1246–1255. URL https://pubs.asha.org/doi/abs/10.1044/1092-4388%282010/09-0121%29.

Liss, J. M., L. White, S. L. Mattys, K. Lansford, A. J. Lotto, S. M. Spitzer et J. N. Caviness. 2009, «Quantifying speech rhythm abnormalities in the dysarthrias», *Journal of Speech, Language, and Hearing Research*, vol. 52, nº 5, doi :10.1044/1092-4388(2009/08-0208), p. 1334–1352. URL https://pubs.asha.org/doi/abs/10.1044/1092-4388%282009/08-0208%29.

Logemann, J. A., H. B. Fisher, B. Boshes et E. R. Blonsky. 1978, «Frequency and cooccurrence of vocal tract dysfunctions in the speech of a large sample of parkinson patients», *Journal of Speech and Hearing Disorders*, vol. 43, nº 1, doi :10.1044/jshd.4301.47, p. 47–57. URL https://pubs.asha.org/doi/abs/10.1044/jshd.4301.47.

Lopez-Moreno, I., J. Gonzalez-Dominguez, O. Plchot, D. Martinez, J. Gonzalez-Rodriguez et P. Moreno. 2014, «Automatic language identification using deep neural networks», dans *2014 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, IEEE, p. 5337–5341.

MacIntyre, A. D., C. Q. Cai et S. K. Scott. 2022, «Pushing the envelope: Evaluating speech rhythm with different envelope extraction techniques», *The Journal of the Acoustical Society of America*, vol. 151, nº 3, doi :10.1121/10.0009844, p. 2002–2026. URL https://doi.org/10.1121/10.0009844.

Maffia, M., R. De Micco, M. Pettorino, M. Siciliano, A. Tessitore et A. De Meo. 2021, «Speech rhythm variation in early-stage parkinson's disease: A study on different speaking tasks», *Frontiers in Psychology*, vol. 12, doi :10.3389/fpsyg.2021.668291,

ISSN 1664-1078. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2021.668291.

Magne, C., M. Aramaki, C. Astésano, R. L. Gordon, S. Ystad, S. Farner, R. Kronland-Martinet et M. Besson. 2005, «Comparison of rhythmic processing in language and music: An interdisciplinary approach», *Journal of Music and Meaning*, vol. 3, p. sec. 5.

Magne, C., C. Astésano, M. Aramaki, S. Ystad, R. Kronland-Martinet et M. Besson. 2007, «Influence of Syllabic Lengthening on Semantic Processing in Spoken French: Behavioral and Electrophysiological Evidence», *Cerebral Cortex*, vol. 17, n° 11, doi : 10.1093/cercor/bhl174, p. 2659–2668, ISSN 1047-3211.

Mairano, P. et A. Romano. 2010, «Un confronto tra diverse metriche ritmiche usando correlatore», *La dimensione temporale del parlato*, vol. 5.

Marczyk, A., B. O'Brien, P. Tremblay, V. Woisard et A. Ghio. 2022, «Correlates of vowel clarity in the spectrotemporal modulation domain: Application to speech impairment evaluation», *The Journal of the Acoustical Society of America*, vol. 152, n° 5, doi : 10.1121/10.0015024, p. 2675–2691.

de Mareüil, P. B., A. Rilliard et A. Allauzen. 2012, «A diachronic study of initial stress and other prosodic features in the french news announcer style: corpus-based measurements and perceptual experiments», *Language and Speech*, vol. 55, n° 2, p. 263–293.

Martin, P. 1981, «Pour une théorie de l'intonation – l'intonation est-elle une structure congruente à la syntaxe ?», dans Rossi et collab. (1981), p. 234–271.

Metter, E. J. et W. R. Hanson. 1986, «Clinical and acoustical variability in hypokinetic dysarthria.», *Journal of communication disorders*, vol. 19, n° 5, p. 347–66.

Meynadier, Y., B. Lagrue, P. Mignard et F. Viallet. 1999, «Effects of l-dopa treatment on the production and perception of parkinsonian vocal intonation», *Parkinsonism and Related Disorders*, vol. 5.

Michelas, A. et M. D'Imperio. 2010a, «Accentual phrase boundaries and lexical access in french», dans *Speech Prosody*, p. 4.

Michelas, A. et M. D'Imperio. 2010b, «Durational cues and prosodic phrasing in french: evidence for the intermediate phrase», dans *Speech Prosody*, p. 4.

Miguel Alonso, B. D. et G. Richard. 2004, «Tempo and beat estimation of musical signals», dans *International Conference on Music Information Retrieval*, Barcelona : Audiovisual Institute, Pompeu Fabra University, p. 158–163.

Méndez, R. et C. Astésano. 2017, «Perception of the downstepped final accent in french», dans *Phonetics and Phonology in Europe (PaPE) 2017, Köln*, p. 102–103.

Narayanan, S. et D. Wang. 2005, «Speech rate estimation via temporal correlation and selected sub-band correlation», dans *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'05).*, vol. 1, IEEE, p. 413–416.

Nazzi, T., J. Bertoncini et J. Mehler. 1998, «Language discrimination by newborns: toward an understanding of the role of rhythm.», *Journal of Experimental Psychology : Human perception and performance*, vol. 24, nº 3, p. 756.

Nespor, M. et I. Vogel. 1989, «On clashes and lapses», *Phonology*, vol. 6, nº 1, doi : 10.1017/S0952675700000956, p. 69–116.

Nocaudie, O., C. Astésano, A. Ghio, M. Lalain et V. Woisard. 2018, «Evaluation de la compréhensibilité et conservation des fonctions prosodiques en perception de la parole de patients post traitement de cancers de la cavité buccale et du pharynx», dans *Proc. XXXIIe Journées d'Études sur la Parole*, p. 196–204, doi :10.21437/JEP.2018-23. URL http://dx.doi.org/10.21437/JEP.2018-23.

Novotný, M., J. Rusz, R. Čmejla et E. Růžička. 2014, «Automatic evaluation of articulatory disorders in parkinson's disease», *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, nº 9, doi :10.1109/TASLP.2014.2329734, p. 1366–1378.

Pampalk, E., A. Rauber et D. Merkl. 2002, «Content-based organization and visualization of music archives», dans *Proceedings of the Tenth ACM International Conference on Multimedia*, MULTIMEDIA '02, Association for Computing Machinery, New York, NY, USA, ISBN 158113620X, p. 570–579, doi :10.1145/641007.641121. URL https://doi.org/10.1145/641007.641121.

Parmenter, C. E. et A. V. Blanc. 1933, «An experimental study of accent in french and english», *PMLA*, vol. 48, nº 2, p. 598–607.

Parsa, V. et D. G. Jamieson. 1999, «A comparison of high precision f0 extraction algorithms for sustained vowels.», *Journal of speech, language, and hearing research : JSLHR*, vol. 42 1, p. 112–126.

Pasdeloup, V. 1990, *Modèle de règles rythmiques du français appliqué à la synthèse de la parole*, thèse de doctorat, Aix-Marseille 1.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot et E. Duchesnay. 2011, «Scikit-learn: Machine learning in Python», *Journal of Machine Learning Research*, vol. 12, p. 2825–2830.

Pellegrino, F. 1998, *Une approche phonétique en identification automatique des langues : la modélisation acoustique des systèmes vocaliques* , Thèse de doctorat, Université Paul Sabatier, Toulouse, France. URL http://www.ddl.cnrs.fr/fulltext/pellegrino/Pellegrino_1998_PhD.pdf.

Pellegrino, F., J.-H. Chauchat, R. Rakotomalala et J. Farinas. 2002, «Can automatically extracted rhythmic units discriminate among languages?», dans *Speech Prosody 2002, Aix-en-provence, France, 11/04/02-13/04/02*, ISCA, p. 563–566. Congres : http ://www.lpl.univ-aix.fr/sp2002/.

Pellegrino, F., J. Farinas et J.-L. Rouas. 2004, «Automatic estimation of speaking rate in multilingual spontaneous speech», dans *Speech Prosody (2004)*, p. 517–520.

Pierrehumbert, J. B. 1980, *The phonology and phonetics of English intonation*, thèse de doctorat, Cambridge : Massachusetts Institute of Technology.

Pike, K. L. 1945, *The Intonation of American English.*, University of Michigan Press.

Pinho, P., L. Monteiro, M. F. d. P. Soares, L. Tourinho, A. Melo et A. C. Nóbrega. 2018, «Impact of levodopa treatment in the voice pattern of parkinson's disease patients: a systematic review and meta-analysis», dans *CoDAS*, vol. 30, SciELO Brasil.

Pommée, T., M. Balaguer, J. Mauclair, J. Pinquier et V. Woisard. 2022, «Intelligibility and comprehensibility: A delphi consensus study», *International Journal of Language & Communication Disorders*, vol. 57, n° 1, doi :https://doi.org/10.1111/1460-6984.12672, p. 21–41. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/1460-6984.12672`.

Post, B. 2003, «Evidence for a constraint-based account of french phrasing and accentuation in different speaking styles», dans *Actes de International Congress of Phonetic Sciences*, p. 1309–1312.

Prince, A. 1983, «Relating to the grid', linguistic inquiry14, 19-100», *Prince1914Linguistic Inquiry1983*.

Prsir, T., J.-P. Goldman et A. Auchlin. 2014, «Prosodic features of situational variation across nine speaking styles in french», *Journal of Speech Sciences*, vol. 4, n° 1, p. 41–60.

Ramus, F. et J. Mehler. 1999, «Language identification with suprasegmental cues: A study based on speech resynthesis», *The Journal of the Acoustical Society of America*, vol. 105, n° 1, p. 512–521.

Ramus, F., M. Nespor et J. Mehler. 1999, «Correlates of linguistic rhythm in the speech signal», *Cognition*, vol. 73, n° 3, p. 265–292.

Ross, M., H. Shaffer, A. Cohen, R. Freudberg et H. Manley. 1974, «Average magnitude difference function pitch extractor», *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 22, n° 5, doi :10.1109/TASSP.1974.1162598, p. 353–362, ISSN 0096-3518.

Rossi, M. 1980, «Le français, langue sans accent ?», *Studia Phonetica Montréal*, vol. 15, p. 13–51.

Rossi, M., A. Di Cristo, D. Hirst, P. Martin et Y. Nishinuma. 1981, *L'Intonation de l'acoustique à la sémantique*, Etudes linguistiques, Klincksieck, ISBN 9782865630103.

Rouas, J.-L., J. Farinas, F. Pellegrino et R. André-Obrecht. 2005, «Rhythmic unit extraction and modelling for automatic language identification», *Speech Communication*, vol. 47, n° 4, p. 436–456. URL `http://www.sciencedirect.com/science/article/B6V1C-4G94G8D-1/2/858e586329f1f2fe156c76a54d67ed32`, ?OLDEditeur(Speech Communication, Elsevier).

Rusz, J., R. Cmejla, H. Ruzickova et E. Ruzicka. 2011, «Quantitative acoustic measurements for characterization of speech and voice disorders in early untreated

parkinson's disease», *The Journal of the Acoustical Society of America*, vol. 129, nº 1, doi :10.1121/1.3514381, p. 350–367. URL https://doi.org/10.1121/1.3514381.

Rusz, J., J. Hlavnička, T. Tykalová, J. Bušková, O. Ulmanová, E. Růžička et K. Šonka. 2016, «Quantitative assessment of motor speech abnormalities in idiopathic rapid eye movement sleep behaviour disorder», *Sleep Medicine*, vol. 19, doi :https://doi.org/10.1016/j.sleep.2015.07.030, p. 141–147, ISSN 1389-9457. URL https://www.sciencedirect.com/science/article/pii/S1389945715009296.

Scheirer, E. et M. Slaney. 1997, «Construction and evaluation of a robust multifeature speech/music discriminator», dans *1997 IEEE international conference on acoustics, speech, and signal processing*, vol. 2, IEEE, p. 1331–1334.

Scheirer, E. D. 1998, «Tempo and beat analysis of acoustic musical signals», *The Journal of the Acoustical Society of America*, vol. 103, nº 1, doi :10.1121/1.421129, p. 588–601. URL https://doi.org/10.1121/1.421129.

Schweitzer, C. et C. Dodane. 2020, «Description de l'accent en français : des premiers grammairiens aux premiers phonéticiens (xvie-début du xxe siècles)», dans *SHS Web of Conferences*, vol. 78, EDP Sciences, p. 09003.

Selkirk, E. O. 1984, *Phonology and syntax: the relationship between sound and structure*, MIT press.

Shattuck-Hufnagel, S. et A. E. Turk. 1996, «A prosody tutorial for investigators of auditory sentence processing», *Journal of psycholinguistic research*, vol. 25, nº 2, p. 193–247.

Sheft, S., M. Ardoint et C. Lorenzi. 2008, «Speech identification based on temporal fine structure cues», *The Journal of the Acoustical Society of America*, vol. 124, nº 1, doi :10.1121/1.2918540, p. 562–575. URL https://doi.org/10.1121/1.2918540.

Sheft, S., V. Shafiro, C. Lorenzi, R. McMullen et C. Farrell. 2012, «Effects of age and hearing loss on the relationship between discrimination of stochastic frequency modulation and speech perception», *Ear and hearing*, vol. 33, nº 6, p. 709–720.

Simon, A.-C. 2020, «Les rythmes dans le slam», *Langage et société*, vol. 171, nº 3, p. 139–169.

Simon, A.-C., A. Auchlin, M. Avanzi et J.-P. Goldman. 2010, «Les phonostyles : une description prosodique des styles de parole en français», dans *Les voix des Français. En parlant, en écrivant*, édité par M. Abecassi et G. Ledegen, Peter Lang : Berne, p. 71–88.

Singh, N. C. et F. E. Theunissen. 2003, «Modulation spectra of natural sounds and ethological theories of auditory processing», *The Journal of the Acoustical Society of America*, vol. 114, nº 6, doi :10.1121/1.1624067, p. 3394–3411. URL https://doi.org/10.1121/1.1624067.

Skodda, S. 2015, «Steadiness of syllable repetition in early motor stages of parkinson's disease», *Biomedical Signal Processing and Control*, vol. 17, doi :https://doi.org/10.1016/j.bspc.2014.04.009, p. 55–59, ISSN 1746-8094. URL https://www.sciencedirect.com/science/article/pii/S1746809414000780, mAVEBA 2013.

Smith, C. L. 2009, «Naïve listeners' perceptions of french prosody compared to the predictions of theoretical models», dans *Proceedings of the third symposium Prosody/discourse interfaces, Paris, September 2009*, Citeseer.

Snyder, D., D. Garcia-Romero, A. McCree, G. Sell, D. Povey et S. Khudanpur. 2018, «Spoken language recognition using x-vectors.», dans *Odyssey*, p. 105–111.

Talkin, D. et W. B. Kleijn. 1995, «A robust algorithm for pitch tracking (rapt)», dans *Speech coding and synthesis*, vol. 495, p. 518.

Te Rietmolen, N. 2019, *Neural signature of metrical stress processing in French*, thèse de doctorat, Université Toulouse le Mirail-Toulouse II.

Tibshirani, R. 1996, «Regression shrinkage and selection via the lasso», *Journal of the Royal Statistical Society : Series B (Methodological)*, vol. 58, n° 1, p. 267–288.

Tilsen, S. et A. Arvaniti. 2013, «Speech rhythm analysis with decomposition of the amplitude envelope: Characterizing rhythmic patterns within and across languages», *The Journal of the Acoustical Society of America*, vol. 134, doi :10.1121/1.4807565, p. 628–39.

Tilsen, S. et K. Johnson. 2008, «Low-frequency fourier analysis of speech rhythm», *The Journal of the Acoustical Society of America*, vol. 124, n° 2, doi :10.1121/1.2947626, p. EL34–EL39. URL `https://doi.org/10.1121/1.2947626`.

Togeby, K. 1965, «Structure immanente de la langue française», *Larousse*.

Troubetzkoy, N., L. Prieto et J. Cantineau. 1939, *Principes de phonologie*, Librairie Klincksieck - Serie Linguistique, Klincksieck, ISBN 9782252034972.

Tykalova, T., M. Novotny, E. Ruzicka, P. Dusek et J. Rusz. 2022, «Short-term effect of dopaminergic medication on speech in early-stage parkinson's disease», *npj Parkinson's Disease*, vol. 8, n° 1, p. 1–6.

Tzanetakis, G. et P. Cook. 2002, «Musical genre classification of audio signals», *IEEE Transactions on Speech and Audio Processing*, vol. 10, n° 5, doi :10.1109/TSA.2002.800560, p. 293–302.

Vaissière, J. 1991, «Rhythm, accentuation and final lengthening», *Music, language, speech and brain*, vol. 59, p. 108–120.

Varnet, L., M. Ortiz-Barajas, R. Guevara Erra, J. Gervain et C. Lorenzi. 2017, «A cross-linguistic study of speech modulation spectra», *The Journal of the Acoustical Society of America*, vol. 141, doi :10.1121/1.4988079, p. 3701–3702.

Vaysse, R., C. Astésano et J. Farinas. 2022a, «Performance analysis of various fundamental frequency estimation algorithms in the context of pathological speech», *The Journal of the Acoustical Society of America*, vol. 152, n° 5, doi :10.1121/10.0015143, p. 3091–3101. URL `https://doi.org/10.1121/10.0015143`.

Vaysse, R., J. Farinas, C. Astésano et R. André-Obrecht. 2021, «Automatic Extraction of Speech Rhythm Descriptors for Speech Intelligibility Assessment in the Context of Head and Neck Cancers», dans *Proc. Interspeech 2021*, p. 1912–1916, doi :10.21437/Interspeech.2021-1736.

Vaysse, R., A. Ghio, C. Astésano, J. Farinas et F. Viallet. 2022b, «Analyse macroscopique des variations et modulations de F0 en lecture dans la maladie de Parkinson : données sur 320 locuteurs», dans *34e Journées d'Études sur la Parole (JEP 2022)*, Association Française de la Communication Parlée, Noirmoutier, France, p. 307–315. URL https://hal.archives-ouvertes.fr/hal-03726999.

Verluyten, S. P. 1984, *Recherches sur la prosodie et la métrique du français*, Universitaire Instelling Antwerpen.

Viallet, F., Y. Meynadier, B. Lagrue, P. Mignard et R. Gantcheva. 2000, «The reductions of tonal range and of average pitch during speech production in off parkinsonians are restored by l-dopa», dans *6th International Congress of Parkinson's Disease and Movement Disorders*, doi :10.13140/RG.2.1.3857.8169.

Wagner, P. 2010, «A time-delay approach to speech rhythm visualization, modeling and measurement», dans *Prosodic Universals comparative studies in rhythmic modeling and rhythm typology*, Rome : Aracne, p. 117–146.

Wang, D. et S. S. Narayanan. 2007, «Robust speech rate estimation for spontaneous speech», *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, n° 8, p. 2190–2201.

Welby, P. S. 2003, *The slaying of Lady Mondegreen, being a study of French tonal association and alignment and their role in speech segmentation*, thèse de doctorat, The Ohio State University.

Wenk, B. J. et F. Wioland. 1982, «Is french really syllable-timed?», *Journal of Phonetics*, vol. 10, n° 2, doi :https://doi.org/10.1016/S0095-4470(19)30957-X, p. 193–216, ISSN 0095-4470.

Woisard, V., C. Astésano, M. Balaguer, J. Farinas, C. Fredouille, P. Gaillard, A. Ghio, L. Giusti, I. Laaridh, M. Lalain, B. Lepage, J. Mauclair, O. Nocaudie, J. Pinquier, G. Pouchoulin, M. Puech, D. Robert et V. Roger. 2021, «C2SI corpus: a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers», *Language Resources and Evaluation*, vol. 55, n° 1, doi :10.1007/s110579-020-09496-3, p. 173–190. URL https://hal.science/hal-02921918.

Woisard, V. et B. Lepage. 2010, «Perception of speech disorders: Difference between the degree of intelligibility and the degree of severity», *Audiological Medicine*, vol. 8, n° 4, doi :10.3109/1651386X.2010.525375, p. 171–178. URL https://doi.org/10.3109/1651386X.2010.525375.

Résumé

La prosodie est un élément essentiel de la parole. Elle constitue un moyen de transmettre l'emphase, le sens, la structure du discours ou encore les émotions. L'un des buts principaux de la prosodie est de segmenter les énoncés de parole en unités linguistiques plus courtes et de les organiser de manière cohérente pour l'auditeur. Les trois principes organisateurs de la prosodie sont : l'intonation, l'accentuation et le rythme. Le rythme de la parole peut être défini comme la récurrence de syllabes accentuées et leur organisation temporelle par rapport aux syllabes inaccentuées. Il joue un rôle primordial dans la structuration temporelle du flot de parole du point de vue du locuteur, et participe également à faciliter la compréhension du message pour l'auditeur. Le rythme est donc un élément central dans l'étude de la prosodie.

Dans cette thèse, nous nous sommes intéressés à l'impact que certaines pathologies peuvent avoir sur la production du rythme de la parole. Plus particulièrement, nous avons étudié deux types de pathologies : la maladie de Parkinson, ainsi que les patients atteints d'un cancer de la cavité buccale ou de l'oropharynx ayant subi un traitement médical. Notre objectif principal a été de proposer une modélisation automatique du rythme de la parole pathologique. Grâce à cette modélisation, nous avons voulu mettre en évidence les régularités rythmiques à différents niveaux prosodiques, dans le but de pouvoir caractériser les stratégies de production de parole misent en jeu chez des personnes atteintes de ces deux pathologies.

Après avoir posé le cadre théorique du rythme dans lequel nous nous plaçons, nous avons pû réaliser un état de l'art des différentes modélisations automatiques du rythmes existantes. Parmi les modélisations automatiques étudiées, nous avons sélectionné celles dont l'implémentation se rapproche au mieux de nos présupposés théoriques. Nous avons alors testé ces méthodes sur un corpus de slam dans le but de sélectionner les méthodologies qui modélisent au mieux la hiérarchie rythmique de la parole. La modélisation que nous avons retenu se base sur l'analyse des modulations lentes (inférieures à 10 Hz) de l'amplitude du signal de parole. Cette méthode appelée le spectre de modulation d'enveloppe (EMS) permet de caractériser la stratégie de segmentation de la parole des locuteurs. Ainsi, nous avons pu observer dans notre corpus de parole pathologique que les personnes présentant de forts troubles de l'articulation des syllabes ont tendance à favoriser une structuration prosodique très régulière. Au contraire, une personne sans troubles apparents de l'articulation présente une structuration prosodique moins régulière. Nous supposons donc que les patients dont l'intelligibilité est faible à cause de troubles articulatoires se focalisent davantage sur une structuration très régulière de leur parole avec des durées de groupes de mots de longueurs équivalentes.

Nous avons par la suite modélisé l'intelligibilité des patients en nous focalisant uniquement sur des indices purement rythmiques issus de l'EMS. Cependant, après

analyse des résultats, les indices rythmiques les plus corrélés au score d'intelligibilité de référence estimés par des médecins ORL étaient en réalité fortement dépendants du débit de parole. Nous avons donc proposé de nouvelles caractéristiques du rythme indépendantes du débit de parole. A l'aide de ces nouveaux paramètres, nous avons pu proposer une représentation en deux dimensions de notre corpus de parole pathologique. Cette représentation basée sur les niveaux principaux de régularités de l'EMS nous a permis de caractériser caractériser et de regrouper les personnes avec des stratégies de segmentation de la parole particulières. L'EMS est donc une modélisation pertinente du rythme de la parole qui permet de caractériser efficacement le rythme de la parole au travers d'une représentation de la régularité des niveaux prosodiques à différents niveaux de hiérarchie.

Mots-clés: Prosodie, Parole pathologique, Rythme, Modélisation automatique, Modulations d'amplitude.

Abstract

Prosody is an essential element of speech. It is a means of conveying emphasis, meaning, speech structure, or emotion. One of the main purposes of prosody is to segment speech utterances into shorter linguistic units and organize them in a coherent way for the listener. The three organizing principles of prosody are : intonation, stress and rhythm. Speech rhythm can be defined as the recurrence of stressed syllables and their temporal organization in relation to unstressed syllables. It plays a key role in the temporal structuring of the speech stream from the speaker's point of view, and also helps to facilitate the comprehension of the message for the listener. Rhythm is therefore a central element in the study of prosody.

In this thesis, we were interested in the impact that certain pathologies can have on the production of speech rhythm. More specifically, we studied two types of pathologies : Parkinson's disease, and patients with cancer of the oral cavity or oropharynx who have undergone medical treatment. Our main objective was to propose an automatic modeling of the pathological speech rhythm. Thanks to this modeling, we wanted to highlight the rhythmic regularities at different prosodic levels, in order to characterize the speech production strategies used by people suffering from these two pathologies.

After having established the theoretical framework of rhythm in which we place ourselves, we were able to carry out a state of the art of the various existing automatic models of rhythm. Among the studied automatic models, we have selected those whose implementation is the closest to our theoretical presuppositions. We then tested these methods on a slam corpus in order to select the methodologies that best model the rhythmic hierarchy of speech. The modeling we have chosen is based on the analysis of slow modulations (lower than 10 Hz) of the speech signal amplitude. This method, called the Envelope Modulation Spectrum (EMS), allows us to characterize

the segmentation strategy of the speakers' speech. Thus, we observed in our corpus of pathological speech that people with strong disorders of syllable articulation tend to favor a very regular prosodic structuring. On the contrary, a person with no apparent articulation disorders presents a less regular prosodic structuring. We therefore assume that patients with poor intelligibility due to articulation disorders focus more on a very regular structuring of their speech with word group durations of equivalent lengths.

We then modeled the patients' intelligibility by focusing only on purely rhythmic cues from the EMS. However, after analysis of the results, the rhythmic indices most correlated with the reference intelligibility score estimated by speech therapist were in fact strongly dependent on the speech rate. We therefore proposed new rhythmic features that are independent of speech rate. Using these new parameters, we were able to propose a two-dimensional representation of our pathological speech corpus. This representation based on the main levels of regularities of the EMS allowed us to characterize and group individuals with particular speech segmentation strategies. The EMS is thus a relevant modeling of speech rhythm that allows us to effectively characterize speech rhythm through a representation of the regularity of prosodic levels at different levels of hierarchy.

Keywords: Prosody, Pathological speech, Rhythm, Automatic modelling, Amplitude modulations.