Diagnostiquer la parole : caractérisation et modélisation automatique

Julie Mauclair

UNIVERSITÉ TOULOUSE III PAUL SABATIER

Diagnostiquer la parole :

Caractérisation et modélisation automatique

MÉMOIRE

soutenu le 9 décembre 2024

pour l'obtention d'une

Habilitation à Diriger les Recherches
Université Toulouse III - Paul Sabatier
(mention informatique)

par

Julie Mauclair

Composition du jury

Rapporteurs :	Elmar Nöth - Professeur, Université d'Erlangen (Allemagne)
	Marie Tahon - Professeur, Université du Mans
	Nathalie Vallée - Directrice de Recherches, Université de Grenoble
Examinateurs :	Véronique Delvaux - Chercheuse FNRS, Université de Mons (Belgique)
	Virginie Woisard-Bassols - Praticien Hospitalier Professeur Associé, Université de Toulouse
	Julien Pinquier - Professeur, Université de Toulouse

Institut de Recherche en Informatique de Toulouse — UMR 5505

Mis en page avec la classe thesul.

SOMMAIRE

INTRODUCTION

La parole, la faculté que l'humain a de s'exprimer est le moyen de communication entre les personnes le plus naturel. Conserver ce lien est essentiel tout au long de notre vie, quelles que soient les conditions environnementales externes ou quel que soit comment nous vieillissons ou de par les pathologies vocales que nous pourrions subir. Il en va de même quand nous apprenons une nouvelle langue et que nous ne voulons pas que notre accent nuise à la communication. Ne pas pouvoir communiquer est un vrai frein à la qualité de vie et peut même aller jusqu'à l'exclusion sociale.

Le traitement automatique de la parole (TAP) représente l'une des avancées technologiques les plus significatives des dernières décennies. Il regroupe plusieurs disciplines dont l'objectif est l'enregistrement, la transmission, la caractérisation et la synthèse de la parole. Les applications en TAP sont vastes et diversifiées : la reconnaissance de la parole, la synthèse de la parole, les assistants vocaux intégrés dans les smartphones, les dispositifs domestiques, la traduction parole à parole... Cependant, les personnes ayant des troubles de la communication sont maintenant davantage marginalisées par cette nouvelle vague de technologies de la parole qui s'intègrent de plus en plus dans la vie quotidienne, mais qui ne sont pas robustes face à la parole atypique, voire pathologique.

Pour mieux tenter de comprendre les défis que peuvent représenter les troubles de la communication sur les systèmes de TAP, il faut déjà comprendre que la parole est un continuum allant de la parole « normale » ou « saine » à la parole « pathologique » en passant par la parole « atypique » . Tous ces qualificatifs caractérisent les différentes variations de la parole et vont être clarifiés dans la suite de ce document. En effet, le fil rouge de ce manuscrit est l'exploration des conséquences des différentes variations de la parole sur les systèmes de TAP, en se concentrant sur des exemples qui se veulent représentatifs des trois déclinaisons principales : la parole saine, la parole atypique et la parole pathologique. Ce continuum correspond aussi finalement à l'historique du TAP. Les premières innovations technologiques ont été effectuées sur de la parole d'homme lue, sans accent, dans des conditions d'enregistrement propres. Puis les premiers défis ont été d'étudier des corpus de parole journalistique comme le corpus ESTER [Gravier 2004], où de la parole semi-préparée a engendré les premières nécessités d'adaptation des systèmes existants. Ensuite, en fonction des appétences des chercheurs et des consortiums formés, des difficultés sont apparues pour traiter des corpus plus variés comprenant des langues peu dotées, des enregistrements de personnes apprenant une langue étrangère, des besoins de transcriptions de réunions en contexte bruité... Et plus récemment, les performances sur ces premiers champs s'améliorant, il est devenu envisageable de s'attaquer à de la parole plus altérée qu'est la parole pathologique.

Ce même continuum représente aussi mon propre parcours en recherche. L'étude de la parole saine constitue le socle de la reconnaissance automatique de la parole (RAP), un des domaines principaux du TAP. La RAP dans un premier temps s'est focalisée sur les caractéristiques acoustiques et linguistiques de la parole produite par des locuteurs sans pathologie vocale voire sans déviance du tout. L'acquisition, le traitement et la modélisation des données de parole saine sont essentiels pour développer des systèmes robustes et fiables de reconnaissance vocale. Cependant, une simple variation de type accent régional, vieillissement de la voix, ou encore altération du canal audio peut engendrer une chute de score de reconnaissance alors que la parole testée est saine. Lors de mon doctorat, j'ai regardé à l'intérieur des mécanismes d'un système de RAP pour voir comment le système pouvait lui-même attribuer un score sur les séquences audio qu'il était à même de transcrire plus ou moins fidèlement. Alors que les corpus étaient exclusivement constitués de parole saine, des altérations subtiles telles que l'accent du locuteur ou encore le style de parole comme un reportage avec des interviews de personnes lambda font que les modèles sous-jacents du système ne retrouvaient pas exactement ce sur quoi ils avaient été entraînés. Plus tard, les encadrements de stage de master et de doctorat m'ont fait remarquer les effets que peuvent avoir les altérations de l'environnement sur les systèmes de reconnaissance. Par exemple, la réverbération avec les travaux de Sébastien Ferreira ou la prononciation des apprenants de L2 dans les expériences de Vincent Laborde. Puis, au cœur de mes projets de recherche plus récents et avec les encadrements de Timothy Pommée et Sebastião Quintas se trouve le traitement de la parole pathologique. Les projets TAPAS, Voice4PD ou encore C2SI et RUGBI m'ont permis de m'intéresser à la parole de patients ayant des pathologies vocales en ayant à cœur l'explicabilité de ces systèmes auprès des thérapeutes. Les défis ici sont multiples, allant de la variabilité extrême des manifestations pathologiques à la rareté des données disponibles pour l'entraînement des modèles. C'est pourquoi je m'attacherai non seulement à exposer les systèmes qui ont été développés combinant apprentissage profond et techniques de traitement du signal ainsi que leurs applications, mais également à comment les corpus de parole pathologiques sur lesquels ils sont appris ont été conçus.

Pour traiter de tout ceci, le plan de ce document est constitué comme suit :

— Le premier chapitre est constitué de quelques éléments précis et fondamentaux d'une chaîne de traitement automatique de la parole qui seront repris dans les chapitres suivants. La collecte des données est notamment un moment essentiel car le matériel utilisé, le protocole et les consignes à donner à la personne responsable de l'enregistrement ainsi qu'aux locuteurs sont cruciaux pour une collecte réussie. Nous nous poserons aussi quelques questions sur la phase d'annotation de ces corpus recueillis en prenant pour exemple deux termes que sont l'intelligibilité et la compréhensibilité dont il faut prendre le temps de clarifier la signification avant de demander à des annotateurs humains de créer la référence en s'appuyant sur ceux-ci. Une fois les données enregistrées au mieux et annotées en prenant soin d'éviter les biais, notamment de compréhension des termes à évaluer, il s'agit de prendre en compte les conditions qui peuvent toutefois comporter des difficultés de traitement. Par exemple, de la réverbération dans le canal audio car les conditions de l'environnement n'ont pas pu être contrôlées est un challenge en soi. Mais savoir que le signal comporte telle ou telle altération peut être un gros indice pour le système. Sachant que le signal est réverbéré par exemple, peut-on proposer une chaîne de traitement qui va augmenter les performances de la RAP ? Le système de transcription est-il capable de savoir quelle partie du signal il va retranscrire fidèlement ?

— Ensuite, durant le chapitre 2, nous verrons plusieurs travaux étudiant la question de la parole

saine et/ou standard dans les systèmes de traitement automatique de la parole. Plusieurs défis scientifiques sont déjà présents dans la parole saine et/ou standard. Tout d'abord, qu'est ce que la parole « standard » en TAP ? Nous nous attarderons par exemple sur la question de l'intelligibilité de la parole saine. Est-ce que parole saine ou standard signifie qu'elle peut offrir la mesure étalon pour calculer les performances en TAP ? Des mesures de confiance issues du système seront aussi étudiées pour augmenter le corpus d'apprentissage de celui-ci avec des séquences audio proches des séquences contenues dans ce dernier pour éviter d'amener trop de variabilité au système. Ces mesures seront également utilisées pour de la combinaison de systèmes de transcription. Enfin, la parole de locuteurs non-natifs sera un dernier exemple de défi concernant la parole saine analysé dans ce chapitre.

— Enfin, le chapitre 3 détaille la constitution des corpus dans lesquels j'ai été impliqué. Puis en décrivant d'abord les outils d'évaluation clinique de la parole déjà disponibles, je montrerai en quoi l'automatique peut fournir un outil d'évaluation supplémentaire au clinicien de par la caractérisation puis la modélisation de la parole pathologique. Les corpus concernant plusieurs pathologies comme la paralysie faciale, Parkinson ou encore les cancers ORL seront étudiés dans ce chapitre.

— Le dernier chapitre dressera le bilan de ce manuscrit et offrira quelques perspectives envisagées à la suite des travaux décrits précédemment.

CHAPITRE 1

ASPECTS DU TRAITEMENT AUTOMATIQUE DE LA PAROLE

SÉLECTION DE PUBLICATIONS RELATIVES À CE CHAPITRE :

- T. Pommée, M. Balaguer, **J. Mauclair**, J. Pinquier, V. Woisard : *Intelligibility and comprehensibility : A Delphi consensus study*, International Journal of Language Communication Disorders, 2021
- S. Ferreira, J. Farinas, J. Pinquier, **J. Mauclair**, S. Rabant : *Analyse de l'effet de la réverbération sur la reconnaissance automatique de la parole*, Journées d'Etudes sur la Parole, 2016
- **J. Mauclair** : *Mesures de confiance en traitement automatique de la parole et applications*. Doctorat de l'Université du Maine, décembre 2006.

Sommaire

1.1 Introduction

Dans ce premier chapitre, nous allons discuter de la chaîne de traitement automatique de la parole, une technologie omniprésente dans notre quotidien. Des applications telles que les assistants vocaux, les systèmes de reconnaissance vocale, et les outils de transcription automatique reposent tous sur cette technologie. Le traitement automatique de n'importe quel type de parole commence par l'acquisition de corpus audio de qualité, étape cruciale pour garantir la précision et la fiabilité des systèmes développés. Ces corpus doivent être soigneusement collectés et annotés pour représenter une variété de conditions et de locuteurs, assurant ainsi que les systèmes soient robustes et capables de fonctionner dans des situations écologiques. Ensuite, il est essentiel pour calibrer les systèmes, d'obtenir une vérité terrain faisant notamment appel à des annotateurs humains pour évaluer les performances des systèmes de traitement de la parole de manière fiable. Pour ce faire, un passage obligé est la définition précise des termes qui seront utilisés comme label dans l'annotation. Une fois soulevés ces deux points essentiels, il s'agit de se demander comment les systèmes réagissent, si des événements ont toutefois altéré le corpus audio, notamment sur le canal de transmission. Nous prendrons l'exemple de la réverbération comme élément altérant les performances du système. Le fait de connaître le taux de réverbération d'un signal présent dans le corpus d'apprentissage peut aider à mettre en place des stratégies pour aiguiller par exemple sur le bon système de reconnaissance de la parole. Et enfin, nous verrons que le système lui-même peut calculer des scores de confiance sur ses propres sorties, fournissant une évaluation précise de ses propres performances et permettant ainsi de mieux répondre aux besoins des utilisateurs finaux.

1.2 Architecture classique des systèmes de traitement automatique de la parole

Les systèmes de traitement automatique de la parole (TAP) sont classiquement composés comme indiqué dans la figure 1.1. Cette figure nous permet de voir différentes phases essentielles pour le bon fonctionnement d'un système de TAP. Dans la partie haute, nous avons la phase d'apprentissage. Pour un système traitant de données audio, un protocole d'enregistrement doit avoir été établi pour que la personne en charge de l'enregistrement donne des consignes précises au locuteur, que ce soit au niveau des tâches à lui faire effectuer oralement, ou encore la position que le locuteur doit avoir face au micro (voir paragraphe 1.3.1). Une fois les données recueillies, la phase d'annotation de celles-ci peut commencer, avec aussi un protocole clair sur ce que les annotateurs doivent rechercher dans le signal (voir paragraphe 1.3.2). Ce corpus une fois annoté devient le corpus d'apprentissage du modèle du système de traitement automatique de la parole. La partie basse montre la phase de test où un enregistrement audio inconnu est donné en entrée du système. Après une phase de pré-traitement et d'extraction des paramètres, le modèle du système de TAP est à même de donner la décision attendue pour la tâche à effectuer.

L'apprentissage des modèles utilisés en TAP se fait généralement grâce à une grande quantité de données préalablement annotées. C'est ce corpus de données audio qui est finalement au centre des déclinaisons de types de parole que nous aborderons par la suite. Dans le langage courant les mots « standard » « atypique » ou « pathologique » ont une signification particulière qui est assez proche des significations données dans le domaine de la recherche informatique en parole. Toutefois, du fait même de la construction du système, les enregistrements de parole utilisés dans ce corpus sont l'essence de la terminologie à adopter. La parole « saine » ou « standard » est directement liée à ce qui se trouve dans le

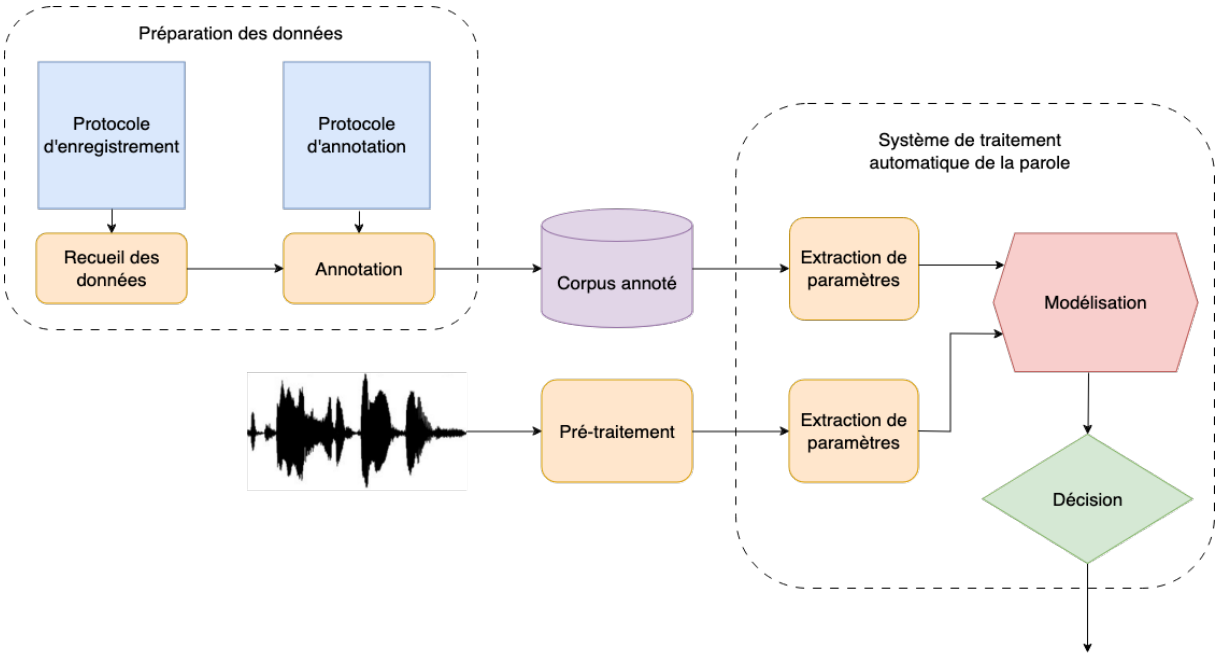

FIGURE 1.1 – Composition d'un système de traitement automatique de la parole

corpus de base puisque c'est ce sur quoi est optimisé le système en termes de performance. Pour le système final, le type de parole qu'il doit savoir caractériser est celui « étalon » qui entre dans la composition de son corpus de base. La préparation des données pour constituer ce corpus revêt alors une importance particulière et fait l'objet du chapitre suivant.

1.3 Préparation des données

1.3.1 Recueil des corpus audio

Dans la majeure partie des projets dans lesquels j'ai été impliquée, je me suis chargée de la mise en place de l'enregistrement audio au niveau matériel et aussi parfois des protocoles à mettre en place pour indiquer les consignes au thérapeute et au locuteur. L'enregistrement de corpus audio pour le traitement automatique de la parole présente plusieurs contraintes. Tout d'abord, le choix des tâches à faire enregistrer est crucial : celles-ci doivent être suffisamment variées pour capturer une large gamme de phénomènes linguistiques et phonétiques, tout en restant représentatives des situations de communication réelles en incluant par exemple une tâche de parole spontanée. Le choix des microphones est également déterminant, car la qualité et la fidélité des enregistrements peuvent considérablement influencer les performances des modèles de reconnaissance vocale ; il est essentiel de sélectionner des microphones qui minimisent le bruit et les distorsions tout en étant pratiques à utiliser dans différents environnements. Le choix s'est souvent porté sur des micros cardioïdes tel que le Rode NT2 ou encore un micro-casque. Une fois les considérations matérielles définies, il faut garder en tête que lorsque l'on travaille avec des patients souffrant de troubles de la parole pathologique, l'inclusion de ces participants nécessite des considérations éthiques et logistiques supplémentaires. Il faut garantir le consentement éclairé, assurer la confidentialité des données,

et parfois adapter les tâches pour être appropriées à leurs capacités. Un protocole clair et détaillé doit être transmis à ceux qui réalisent les enregistrements pour garantir la cohérence et la comparabilité des données recueillies. Ce protocole doit inclure des instructions sur le réglage des équipements, la gestion des variations environnementales, et la manière de guider les participants tout au long des tâches d'enregistrement, afin de standardiser les conditions et réduire les biais potentiels dans les données. De plus, l'implication du Comité de Protection des Personnes (CPP) est indispensable pour toute étude impliquant des participants humains. Ce comité a pour mission de veiller à ce que les droits, la sécurité et le bien-être des participants soient protégés. Il évalue la pertinence du protocole de recherche, vérifie les procédures de consentement et s'assure que les risques encourus par les participants sont minimisés. L'approbation du CPP est donc une étape cruciale pour garantir la conformité éthique des enregistrements de corpus audio, particulièrement lorsqu'ils concernent des populations vulnérables telles que les patients atteints de pathologies de la parole.

1.3.2 S'accorder sur la terminologie

Une fois l'enregistrement effectué et avant de traiter les données, il faut les annoter pour avoir la vérité terrain. Pour cette labellisation, il est nécessaire et primordial de s'accorder sur la terminologie à utiliser. Si des ambivalences persistent entre annotateurs, la vérité terrain sera caduque.

Par exemple, une des études que j'ai encadrée concerne l'ambiguïté particulière résidant entre les termes « intelligibilité » et « compréhensibilité », deux termes fréquemment employés notamment dans l'analyse de la parole pathologique [Pommée 2021b]. Cette ambiguïté est observée à la fois dans l'usage terminologique par les cliniciens, dans les batteries d'évaluation de la parole existantes et dans la littérature scientifique.

Les définitions des termes relatifs à la parole varient en fonction du domaine de pratique ou de recherche de ceux qui les utilisent. Les experts en production ou perception de la parole peuvent avoir des conceptions différentes de l'« intelligibilité ». Cette diversité de sens peut provoquer des problèmes de communication entre professionnels, impactant la pratique clinique, mais aussi la recherche scientifique [Denman 2019, Walsh 2005, Walsh 2006]. En contexte clinique, ces divergences peuvent nuire à l'efficacité de la prise en charge des patients au sein d'une équipe pluridisciplinaire et entraîner des malentendus lors du transfert d'un patient d'un professionnel à un autre. Sur le plan scientifique, elles compliquent les débats entre chercheurs, rendant difficiles la comparaison et la combinaison des résultats de recherche, ce qui limite notamment la répercussion des résultats scientifiques. Ainsi, les avancées tendent à être cloisonnées au sein des différents domaines de recherche, plutôt que de favoriser un progrès global et transdisciplinaire. De plus, l'absence de terminologie consensuelle constitue un obstacle majeur à la communication entre les domaines clinique et scientifique, empêchant le transfert des découvertes fondamentales vers des applications cliniques [Denman 2019, Roulstone 2015].

Face à ce déficit de consensus concernant la terminologie liée à l'évaluation de la parole et de son impact dans les domaines clinique et scientifique, nous avons entrepris une enquête internationale selon la méthode Delphi [Chalmers 2019]. L'objectif de cette étude était de formuler une définition plus exhaustive et consensuelle des deux termes : intelligibilité et compréhensibilité. La méthode Delphi consiste en un processus en trois étapes successives afin d'obtenir un consensus parmi un groupe de participants ayant une expertise dans un domaine spécifique [Birko 2015, Linstone 2002b]. Elle emploie des questionnaires remplis indépendamment par chaque expert. Chaque tour est suivi d'une synthèse des réponses agrégées,

permettant aux participants de réévaluer leurs opinions. Facilement applicable en ligne, cette méthode présente des avantages significatifs tels que son coût modéré et l'élimination des contraintes géographiques [Hartman 1995, Linstone 2002a, Turoff 1996]. Son caractère quasi anonyme est un argument crucial en faveur de cette méthode [Sinha 2011, von der Gracht 2012]. L'identité des participants demeure connue uniquement du modérateur, préservant l'anonymat entre les participants et favorisant une expression libre dépourvue de pressions sociales ou professionnelles. En outre, la nature quasi anonyme combinée à l'utilisation de plusieurs tours consécutifs et aux feed-back structurés fournis aux participants après chaque tour permet de minimiser les biais inhérents au processus de recherche de consensus [Chalmers 2019].

Cette étude Delphi a abouti à une définition étayée de l'intelligibilité et de la compréhensibilité et de leur évaluation, intégrant l'ensemble des éléments consensuels identifiés tout au long du processus. La figure 1.2 résume cette définition et illustre la relation entre les deux concepts.

L'intelligibilité et la compréhensibilité sont deux notions liées à la parole, bien qu'ils diffèrent dans leur signification. Tous deux évaluent les compétences de production du locuteur et jouent un rôle crucial dans la communication. Il est essentiel de noter que, même si l'accent est mis sur la production de la parole, il ne faut pas négliger les facteurs de perception de l'auditeur et le fait que certains auditeurs puissent avoir des troubles auditifs.

L'intelligibilité se concentre sur la reconstruction d'un énoncé au niveau acoustico-phonétique, où l'information dépend du signal acoustique. Cette reconstruction est facilitée par les capacités de production phonétique-acoustique du locuteur et les compétences de décodage acoustico-phonétique de l'auditeur. Perceptuellement, l'intelligibilité est évaluée de manière optimale avec des stimuli à faible prédictibilité tels que les phonèmes, syllabes, pseudo-mots, mots en paires minimales et phrases non prédictibles. Ces phrases permettent une évaluation plus fonctionnelle où le phénomène de coarticulation est pris en compte mais en faisant en sorte de court-circuiter le processus de compensation cognitive de l'auditeur, en faisant fi du contexte sémantique ou linguistique. D'un point de vue instrumental, l'intelligibilité peut être mesurée à l'aide de mesures acoustiques, en s'appuyant sur les consonnes, voyelles, glides produits dans les tâches nommées précédemment. Le débit de parole, accentuation, et qualité de la voix participent également à l'intelligibilité.

La compréhensibilité, en revanche, se réfère à la reconstruction d'un message au niveau sémantique-discursif, après la phase de reconstruction acoutisco-phonétique. Elle implique des éléments contextuels indépendants du signal, comme le contexte linguistique ou non verbal. Bien que l'intelligibilité soit une composante de la compréhensibilité, cette dernière ne dépend pas exclusivement de la précision du décodage acoustico-phonétique. La compréhensibilité concerne l'aspect utile ou encore fonctionnel de la communication et, du point de vue perceptif, elle est mieux appréhendée à travers des évaluations centrées sur le sens. Cela signifie qu'elle prend en considération les processus cognitifs descendants capables de compenser la dégradation des informations acoustiques et phonétiques. Jusqu'à présent, aucune mesure instrumentale dite « objective » n'est spécifiquement conçue pour évaluer la compréhensibilité telle quelle, c'est-à-dire la transmission du sens global du message. Toutefois, certains paramètres suprasegmentaux, tels que les mesures de débit et d'intonation, contribuent objectivement à la compréhensibilité.

Finalement, une façon de différencier l'intelligibilité et la compréhensibilité dans la définition proposée est le terme utilisé pour désigner le processus de (re)construction respectif : le terme « énoncé » est employé pour référer au contenu acoustique et phonétique de la parole (et, par conséquent, à l'intelligibilité), tandis que le terme « message » est utilisé comme un terme plus général référant à la (re)construction au niveau sémantique (c.-à-d. à la compréhensibilité, qui comprend également les éléments d'intelligibilité). En effet,

comme deux participants l'ont suggéré dans leurs commentaires, alors que le terme « message » peut être défini comme faisant référence au « thème ou à l'idée sous-jacente » (donc, avec une idée de « signification »), le terme « énoncé » se détache du contenu sémantique communiqué et se rapporte plutôt au signal acoustique transmis.

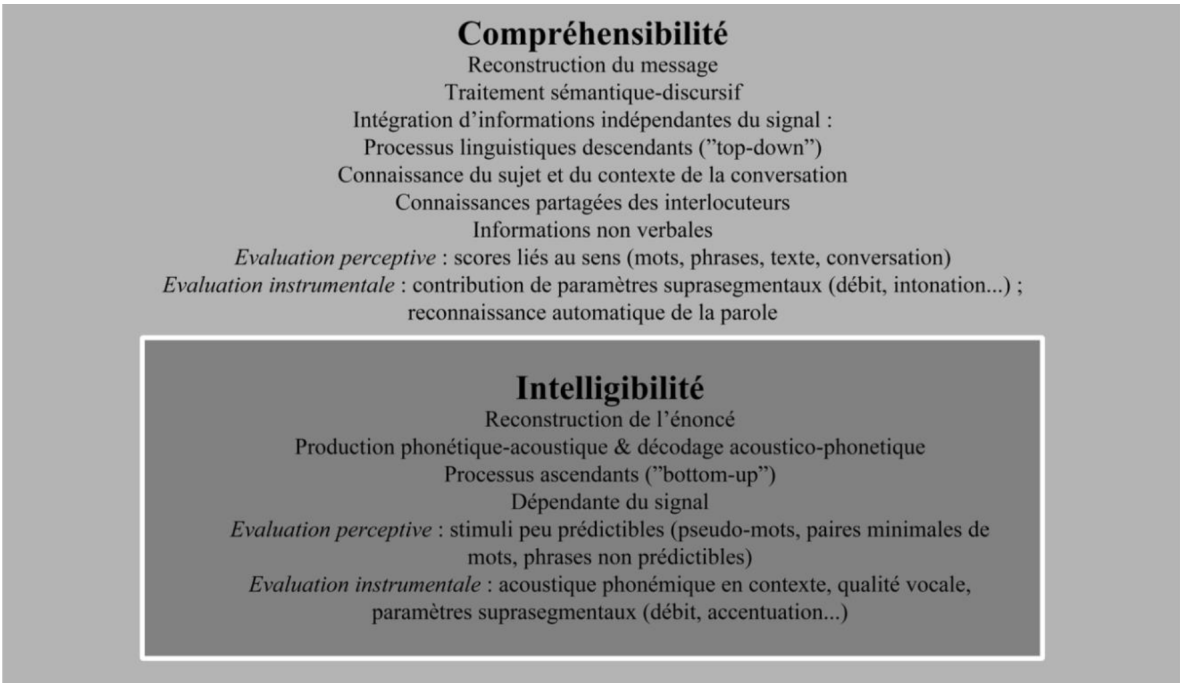

FIGURE 1.2 – L'intelligibilité et la compréhensibilité dans la production de la parole, tirée de [Pommée 2021b]

Cette phase de recherche d'accord sur la terminologie est essentielle à tout bon déroulement d'un projet. Elle est nécessaire à mettre en place pour une phase d'annotation de corpus efficiente. En effet, une fois que la terminologie a été formellement acceptée par tous les intervenants du projet, la phase d'annotation du corpus peut être initiée. Cette étape revêt une importance capitale dans le processus de développement et d'évaluation des systèmes automatiques. L'annotation du corpus permet d'établir une vérité terrain, une référence fiable, à partir de laquelle les performances des systèmes automatisés peuvent être évaluées de manière objective. Cette démarche vise à créer une base de données annotée solide sur laquelle les chercheurs peuvent mesurer avec justesse l'efficacité et la précision de leurs systèmes automatiques.

1.3.3 La phase d'annotation

En traitement automatique, toute hypothèse doit être validée sur des expériences dont la performance est calculée en comparant le résultat à une vérité terrain. La vérité terrain dépend de l'objectif de la tâche à accomplir. Parmi les tâches à annoter manuellement, certaines sont moins chronophages que d'autres, comme l'annotation de zones de parole/non parole ou encore la tâche de segmentation en Parole/Musique/Bruit. La complexité de l'annotation va croissant avec la multiplication ou la superposition des classes qui sont à découvrir, telles que dans la tâche de segmentation et regroupement en locuteurs

ou celle de l'identification en locuteur. Des problèmes de tuilages et donc de superposition peuvent ainsi rendre ce type d'annotation complexe. La transcription orthographique par exemple, demande à un annotateur humain environ 8 fois le temps réel de prononciation. Dans [Bazillon 2008], les auteurs rapportent un rapport de 8 à 9 fois le temps réel de parole sur le corpus ESTER [Gravier 2004]. En fonction de la difficulté de l'annotation (transcription orthographique ou assignation en locuteurs) et du type de parole (spontanée ou préparée), le temps d'annotation augmente comme indiqué sur le tableau 1.1 tiré de [Bazillon 2008].

TABLEAU 1.1 – Durée totale des annotations en transcription orthographique et assignation des différents locuteurs sur de la parole préparée (2h08) et de la parole spontanée (2h10), tableau extrait de [Bazillon 2008]

	Parole préparée (2h08)	Parole spontanée (2h10)
Transcription manuelle	17h36	19h33
Assignation en locuteur	1h17	2h13

La tâche d'annotation manuelle devient d'autant plus complexe et longue à mesure que le niveau de détail requis augmente. Elle l'est également d'autant plus si la qualité du canal de communication est altérée mais aussi en fonction du niveau d' « atypisme » du locuteur (ce terme est repris et défini au chapitre 2). Ces facteurs ainsi que d'autres sont énumérés dans la figure 1.3, adaptée des travaux de Sébastien Ferreira [Ferreira 2021] et affectent également le taux d'erreur mot (WER) d'un système de reconnaissance de la parole (SRAP). Le taux d'erreur de reconnaissance automatique de la parole est étroitement lié à divers facteurs influençant la qualité et la fidélité du processus. Parmi ces éléments, la capture et le stockage audio jouent un rôle majeur en regroupant les variabilités induites par le matériel d'enregistrement et les méthodes de stockage et de compression du signal vocal. Les caractéristiques de l'environnement sonore constituent un autre paramètre significatif, englobant les variabilités acoustiques externes à la parole, comme la réverbération, les bruits ambiants et la superposition de locuteurs, susceptibles d'altérer la clarté du signal. La variabilité inter-locuteur, envisagée principalement d'un point de vue physique et physiologique, c'est-à-dire son âge, son accent, sa pathologie..., ajoute une complexité supplémentaire. L'expression elle, rassemble les variabilités liées au contexte d'énonciation du locuteur en lui-même, à savoir son émotion ressentie, la tâche qu'il est en train de réaliser... Enfin, les aspects lexicaux, comprenant le sujet abordé dans l'enregistrement, les mots connus ou non du SRAP, contribuent également aux défis inhérents à la reconnaissance automatique de la parole. Bien sûr, il est plus facile d'obtenir une bonne transcription quand le signal est de bonne qualité que lorsqu'il est bruité. Il est plus facile également d'obtenir de bons résultats quand le locuteur est d'un âge compris entre 30 et 55 ans, sans accent régional.

Dans la suite de ce chapitre, nous verrons en quoi prendre en compte l'altération de la parole, que ce soit une altération due à l'environnement sonore ou à une voix différente de celles constituant le corpus d'apprentissage d'un SRAP permettent d'extraire du signal une information précieuse et donc par exemple, de prédire la performance d'un système de reconnaissance automatique de la parole. La prise en compte d'une singularité de la parole énoncée permet de comprendre les performances du système de RAP, et d'adapter ou d'utiliser ses résultats pour l'application visée. Le premier exemple qui peut être généralisé à d'autres conditions environnementales est l'étude de la réverbération du signal afin de prédire

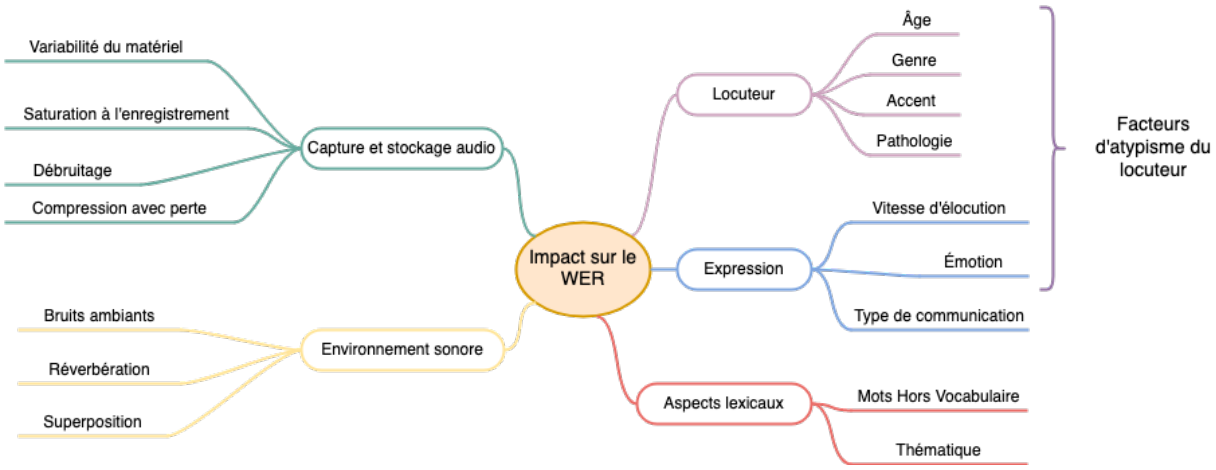

FIGURE 1.3 – Carte heuristique sur les sources de variabilité qui impactent le WER (Word Error Rate ou taux d'erreur mots) des SRAP (systèmes de reconnaissance automatiques de la parole), adaptée de [Ferreira 2021]

les performances d'un SRAP et de rediriger la transcription vers le système le plus performant.

1.4 Comment le système peut-il réagir face aux altérations sur le signal audio ?

Plusieurs facteurs influencent les performances des systèmes de reconnaissance de la parole. Tout au long du canal de transmission peuvent intervenir des artéfacts qui perturbent la robustesse de la reconnaissance. Des erreurs peuvent être liées à une mauvaise captation du signal, à l'environnement sonore, mais aussi au fait que les données d'apprentissage ne soient pas équilibrées ou différentes de l'usage qui en est fait. Comme le système a souvent été appris sur des locuteurs sains (nous verrons dans le chapitre 2 comment on peut qualifier plus précisément ce type de parole), les locuteurs dont la parole s'écarte de la norme, qu'on appellera locuteurs atypiques, sont mal reconnus par le système, ce qui entraîne une dégradation des performances. Si le canal par lequel passe la transmission est lui-même altéré, par du bruit dans l'environnement d'enregistrement ou encore la réverbération de la pièce, le système va aussi être perturbé et les taux de reconnaissance seront dégradés.

Les connaissances sur la parole, sur l'environnement dans lequel elle est produite, nous fournissent des indications essentielles pour la mise en place d'un système de traitement adéquat. En amont de la chaîne de traitement, des stratégies de choix du SRAP peuvent être opérées en fonction des dégradations perçues. Dans son travail de thèse [Ferreira 2021], Sébastien Ferreira propose plusieurs contributions en fonction de trois dégradations majeures, le bruit ambiant, la réverbération et la parole superposée. Ma collaboration avec Sébastien a été sur la partie réverbération et c'est pourquoi je prendrais cet exemple pour illustrer en quoi la prise en compte de la dégradation due à l'environnement peut être d'une grande aide pour améliorer les résultats des systèmes. La section 1.4.1 traite donc l'exemple de la réverbération altérant le signal audio et permettant de choisir le système de transcription le plus adapté. En sortie du système, les pertes de performances peuvent être évaluées par celui-ci pour décider des actions à effectuer

en fonction de l'objectif final.

Ensuite, la section 1.4.2 décrit comment des indices intrinsèques au SRAP peuvent permettre au système de juger de ses propres performances sur les sorties qu'il produit. Ces mesures de confiance peuvent être ensuite utiles dans plusieurs applications, dont des exemples seront traités sur de la parole saine dans le chapitre 2.

1.4.1 Critères *a priori* pour choisir le système de reconnaissance de la parole

L'environnement dans lequel est énoncée la parole peut perturber la performance des systèmes de transcription. La connaissance *a priori* de la qualité de l'environnement permet d'aiguiller l'enregistrement audio vers le système de transcription qui aura les meilleures réalisations. Comme nous l'avons vu précédemment, plusieurs facteurs de l'environnement peuvent affecter la qualité du signal audio comme les bruits ambiants, la réverbération ou encore la parole superposée. Tous ces facteurs sont étudiés dans [Ferreira 2021]. En exemple, nous allons analyser ce que produit sur le signal de parole la réverbération d'une salle.

Dans une étude sur la dégradation engendrée par la réverbération de la pièce [Ferreira 2020a], Sébastien Ferreira analyse l'effet de celle-ci sur les performances d'un système de reconnaissance de la parole (SRAP). La distorsion engendrée par la réverbération est le flou temporel. Le flou temporel provoque un chevauchement du phonème précédent sur le phonème en cours. Sur la figure 1.4, tirée de [Petrick 2008], nous pouvons voir l'énergie résiduelle du phonème précédent qui déteint sur le phonème courant. Maintenant la question est de savoir comment se comportent les systèmes de RAP lorsque les phonèmes se chevauchent.

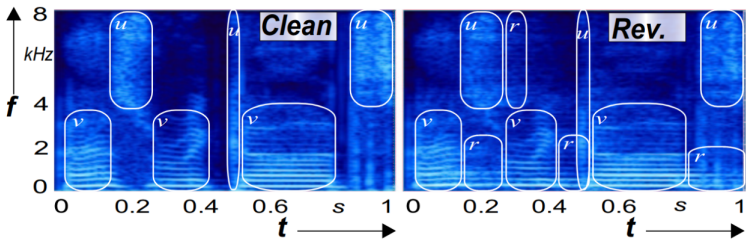

FIGURE 1.4 – Illustration des perturbations causées par la réverbération, extrait de [Petrick 2008]. Ici, la lettre v correspond aux phonèmes voisés, u aux phonèmes non-voisés et r à la réverbération due au flou temporel

Sur le corpus de parole du Wall Street Journal [Paul 1992] , des réponses impulsionnelles de différentes salles provenant du REVERB Challenge [Kinoshita 2013] sont appliquées afin de réverbérer la parole artificiellement. Les erreurs commises par un SRAP sont alors analysées pour vérifier l'influence de la réverbération sur le système. Ce système, basé sur Kaldi [Veselỳ 2013], est un décodeur acoustico-phonétique, s'affranchissant du modèle de langage, afin de prédire directement une suite de phonèmes. En fonction de la taille de la pièce (petite, moyenne ou grande) et de la distance au microphone (proche ou loin), le taux d'erreur phonémique (PER) est altéré plus ou moins par rapport à une condition « propre » sans convolution avec une réponse impulsionnelle comme le montre le tableau 1.2.

Deux facteurs impactent donc fortement les performances des SRAP :

TABLEAU 1.2 – Résultats en termes de PER en fonction des différentes conditions de réverbération : moyenne, écart-type et pourcentages de substitution, insertion, délétion de phonème.

Taille salle	Propre	Petite		Moyenne		Grande	
Distance		proche	loin	proche	loin	proche	loin
PER en %							
Moyenne	9,8	14,1	24,2	28,5	50,6	31,7	63,8
Écart-type	5,6	6,7	8,9	9,2	8,9	9,5	7,4
Ratio des erreurs des phonèmes en %							
Substitution	61,2	59,8	60,8	60,7	56,2	61,1	53,8
Insertion	15,3	15,5	11,4	9,4	4,0	9,2	2,1
Délétion	23,5	24,9	27,8	29,9	39,8	29,7	44,1

TABLEAU 1.3 – Influence de l'âge du locuteur (a) et de l'accent du locuteur (b) sur le taux d'erreur mots (résultats en %)

	Adulte	Senior
Homme	34,6	43,2
Femme	46,8	61,2
Ensemble	36,4	47,8

(a) Âge du locuteur

	Moyenne Accent/ Non-Natif	Natif(US)
SRAP classique	39,4	14,5
SRAP adapté	23,1	12,4

(b) Accent du locuteur

— la taille de la pièce. Plus la taille de la pièce augmente et plus l'énergie provenant de la réverbération est importante, car le temps de réverbération augmente.

— la distance au microphone. Plus la distance au microphone augmente et plus l'énergie de la parole provenant du trajet direct est atténuée. Ces deux facteurs modifient le ratio entre l'énergie de la parole et l'énergie de la réverbération.

Grâce à cette analyse, nous pouvons voir que le flou temporel influe effectivement sur les performances du système de reconnaissance. Des facteurs environnementaux tels que la réverbération mais aussi la parole superposée à de la musique peuvent causer ces baisses de performance. La variabilité intrinsèque du locuteur telle que son âge, son genre ou encore son accent influent également tel que présenté dans les tableaux 1.3a et 1.3b . Ces résultats sont extraits respectivement de [Vipperla 2008] et [Turan 2020].

Ces facteurs propres au locuteur mais aussi l'expression de celui-ci, sa vitesse d'élocution, son émotion et évidemment son type de pathologie peuvent également avoir un impact. Pour avoir une idée des performances possibles d'un SRAP, il devient crucial d'évaluer, d'estimer et de prédire la qualité d'une transcription.

Par exemple, dans le cas de la parole réverbérée en condition mono-canal sans connaissance de la RIR, les auteurs de [Ferreira 2020b] proposent une nouvelle mesure appelée Excitation Behaviour (EB). Cette mesure analyse les résidus de la prédiction linéaire (PL). Sur les fenêtres voisées de la parole, le résidu de la PL contient des informations sur l'instant de fermeture glottale et sur la source d'excitation [Ananthapadmanabha 1979]. Lorsque la parole est réverbérée, la différence entre les impulsions glottales et la source d'excitation est plus faible. C'est cette distorsion des résidus de la PL qui sera exploitée par l'EB.

Afin d'évaluer la performance de la prédiction du WER grâce à l'EB, nous avons calculé l'erreur de prédiction absolue moyenne (MAE pour Mean Average Error) et l'écart-type (SD pour Standard Deviation). Le MAE et le SD sont indiqués pour toutes les conditions de réverbération testées. Les résultats de la prédiction sont présentés dans le tableau 1.4 et comparés à d'autres mesures de l'état de l'art telles que :

— SRMR+ : SRMR (standardized root mean squared residual) et SRMR normalisés [Falk 2010],
— Slope : ici c'est la valeur de « floored ratio of spectral substraction » [Tachioka 2013],
— Neg-side : une méthode de Spectral Decay Distribution qui utilise la variance négative et le skewness [Dumortier 2014],
— LP-kurto : moyenne des kurtosis des résidus de la PL d'ordre 10 [Gillespie 2001].

TABLEAU 1.4 – Resultats de prédiction (Word Error Rate et Phoneme Error Rate) avec une régression Multi Layer Perceptron.

	SRMR+	Slope	Neg-side+	LP-kurto	EB
WER (%)					
MAE	17,75	18,44	17,45	18,33	**13,66**
SD	14,26	14,95	13,75	15,69	**12,63**
PER (%)					
MAE	10,76	12,59	10,82	11,43	**7,86**
SD	8,14	9,04	7,75	9,15	**6,25**

La mesure EB permet donc d'obtenir une prédiction des performances des systèmes de RAP. En fonction des situations, l'élaboration de mesures acoustiques analysant le signal de parole peut donc aider à prédire la qualité des transcriptions des SRAP et donc choisir le meilleur système de transcription *a priori*. En aval d'un chaîne de traitement automatique, l'élaboration de telles mesures peut également avoir lieu au sein du modèle lui-même pour fournir une indication *a priori* sur la performance que va atteindre le système global. Nous allons dans le prochain paragraphe donner un exemple de scores que peut fournir un SRAP pour diagnostiquer son propre comportement.

1.4.2 Auto diagnostic du système de traitement

Atteindre de très bonnes performances avec un système de transcription sur n'importe quelle dégradation de la parole n'est pas une fin en soi. En fonction du choix des données d'apprentissage par rapport à l'application visée, on peut avoir de très bons scores de WER par exemple. Mais parfois, la déviance des sorties du système de traitement de la parole par rapport à une voix « canonique » dont nous verrons une définition dans le chapitre suivant, est déjà une information qui mérite d'être prise en compte. Savoir mesurer et évaluer la distance entre les données de départ et les données applicatives est une bonne façon de déterminer à quel point l'ensemble de test est éloigné du corpus d'apprentissage. En fonction de cette distance et de la finalité, on peut développer des stratégies diverses. En pathologie, la distance à de la parole saine peut permettre au thérapeute d'orienter le patient vers tel ou tel exercice, en apprentissage de la lecture, une application peut proposer de faire répéter ou de faire écouter la bonne prononciation... Cette distance par rapport à « l'attendu » peut être fournie par le système lui-même. Les mesures de confiance (MC) permettent au départ d'évaluer la confiance qu'ont les SRAP sur l'exactitude des hypothèses de transcription fournies. Aujourd'hui ce score a évolué pour indiquer la fiabilité globale des systèmes de

traitement automatique. Une mesure de confiance associée à une hypothèse de reconnaissance est une estimation de la fiabilité de cette hypothèse. La mesure de confiance $MC(h)$ associée à une hypothèse h appartient à l'intervalle $[0, 1]$ et peut être interprétée comme étant la probabilité que l'hypothèse soit correcte ou non. Idéalement, $MC(h)$ vaut 0 si l'hypothèse h est incorrecte, 1 si elle est correcte.

Les mesures de confiance sont utilisées dans de nombreux domaines du traitement de la parole [Lee 2001] comme la reconnaissance de la parole [Wessel 2005, Cox 2002], les systèmes de dialogue [San-Segundo 2001] ou encore l'identification des langues [Metze 2000].

Plus récemment, dans [Ghannay 2015], les auteurs prédisent pour chaque mot, si le mot est correct ou non : le score est binaire. Pour calculer ce score, ils utilisent une architecture de réseau de neurones, qui est chargée d'attribuer un label (correct ou incorrect) pour chaque mot d'une hypothèse de transcription. La classification utilise une représentation de type « word embedding » (plus précisément « word2vec » [Mikolov 2013]) comme entrée, en addition des MC extraites des graphes de décodage du SRAP et des paramètres lexicaux et syntaxiques.

Les mesures de confiance que j'ai proposées et élaborées dans [Mauclair 2006a] sont issues d'un SRAP basé sur CMU Sphinx III [Deléglise 2005] et permettent un auto-diagnostic. Elles étaient de trois types :

— La première (MC_1) est basée sur une mesure acoustique déjà connue qui exploite la distance introduite par l'utilisation d'un modèle de langage et d'un dictionnaire de prononciations entre les scores acoustiques donnés par un modèle acoustique contraint et ce même modèle utilisé sans contrainte.
— La deuxième mesure (MC_2) provient du modèle de langage et prend en compte son comportement de repli (*backoff*) lors du décodage par le SRAP qui a fourni l'hypothèse évaluée.
— La troisième (MC_3) s'appuie sur le calcul de la probabilité *a posteriori* d'un mot issue d'un réseau de confusion. Celui-ci représente l'espace de recherche du SRAP où les mots en compétition au même instant sont alignés. La probabilité *a posteriori* d'un mot regroupe intrinsèquement les scores acoustiques et linguistiques du SRAP.

Ces trois mesures prises séparément apportent une information sur la pertinence d'un mot et permettent par exemple de repérer des zones erronées ou correctes dans l'hypothèse fournie par un SRAP.

La meilleure mesure en terme d'entropie normalisée croisée sur un corpus de développement est une combinaison de le mesure de confiance calculée à partir de la probabilité *a posteriori* d'un mot et de la mesure de confiance estimée en fonction du repli du modèle de langage. Elle sera notée MC_{fusion}.

La figure 1.5 montre la mesure de confiance MC_{fusion} ainsi que les mesures qui la composent pour les comparer en termes de taux de mots émis incorrects en fonction du taux de rejet sur un corpus de développement.

Nous prendrons un exemple dans le prochain chapitre en nous servant de ces mesures pour filtrer les séquences de mots d'un corpus non annoté. Ces séquences de mots seront utilisées pour augmenter le corpus d'apprentissage d'un SRAP.

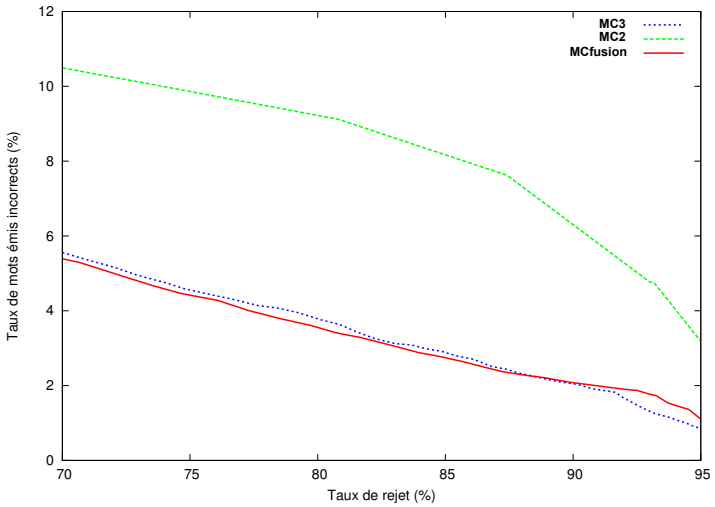

FIGURE 1.5 – Taux de mots émis incorrects en fonction du taux de rejet pour trois mesures de confiance : MC_2, MC_3 et MC_{fusion}

1.5 Conclusion

Les connaissances sur la parole, quelle qu'elle soit, sur l'environnement dans lequel est produite celle-ci, nous fournissent des indications essentielles pour la mise en place d'un système de traitement adéquat. En combinant des méthodes de traitement avancées et des corpus de haute qualité, il est donc possible de concevoir des systèmes capables de répondre aux exigences de précision et de robustesse des applications modernes.

Dans les prochains chapitres, cette connaissance sur la chaîne de traitement peut d'abord commencer par s'étalonner sur de la parole « canonique » et « standard » . A quoi se réfère-t-on exactement ? Qu'est-ce qu'une parole « normale » ? Il est primordial de poser les bases d'une réflexion sur ces termes afin de pouvoir ensuite parler de parole déviante et/ou atypique. Ensuite, quand on s'attarde sur les pathologies vocales, le terme de parole saine apparaît. Une parole saine est-elle forcément normale dans le sens de normée ? Est-elle forcément une parole parfaite pour les applications que l'on veut faire après ?

Le chapitre suivant permet de repérer ces différents concepts et de comprendre l'importance du cadre à mettre en place avant de s'attaquer à mesurer un écart à cette norme. En effet, même si le locuteur n'a pas de trouble vocal, en fonction de la complexité de la tâche que l'on va lui donner à effectuer tout d'abord, le SRAP ne va pas réagir de la même façon. Ensuite, du point de vue perceptif, la parole saine est déjà multiple et si on parle d'intelligibilité, la référence pour l'auditeur va quand même induire des différences en fonction des accents, de l'âge ou de la nationalité du locuteur. Ces altérations dans les prononciations des locuteurs peuvent induire une grande diminution des performances des systèmes. Les mesures de confiance introduites dans le présent chapitre vont alors être une aide pour évaluer quelle partie de l'énoncé du locuteur est susceptible d'être fournie avec succès au système de TAP. Elles peuvent également indiquer les divergences que le système détecte par rapport à son corpus d'apprentissage, entraîné sur de la parole standard.

CHAPITRE 2

QUELLE EST NOTRE RÉFÉRENCE DE PAROLE « TYPIQUE » ? QU'EST CE QU'UNE PAROLE « SAINE » ?

SÉLECTION DE PUBLICATIONS RELATIVES À CE CHAPITRE :

- **J. Mauclair**, Y. Estève, S. Petit-Renaud, P. Deléglise : *Automatic Detection of Well Recognized Words in Automatic Speech Transcriptions*, Language Resources and Evaluation Conference, 2006
- T. Pommée, M. Balaguer, J. Pinquier, **J. Mauclair**, V. Woisard, R. Speyer : *Relationship between phoneme-level spectral acoustics and speech intelligibility in healthy speech : a systematic review*, Speech, Language and Hearing 2021
- B. Lecouteux, G. Linarès, Y. Estève, **J. Mauclair** : *System Combination by Driven Decoding*, ICASSP 2008
- V. Laborde, T. Pellegrini, L. Fontan, **J. Mauclair**, H. Sahraoui, J. Farinas, : *Pronunciation assessment of Japanese learners of French with GOP scores and phonetic information*, INTERSPEECH 2016.

Sommaire

2.1 Introduction

Dans ce chapitre, nous explorerons en tant qu'objet de recherche les défis soulevés par la parole et ses différentes déclinaisons, déjà quand il s'agit de parole sans trouble vocal. En effet, comme les premiers corpus disponibles se sont trouvés être des corpus de personnes sans accent et lisant un texte connu à l'avance, les premiers systèmes ont été appris, évalués et optimisés sur ce type de parole, qui de fait est devenue la parole standard du moins pour les systèmes de traitement automatique de la parole. Du moment où l'on cherche à s'éloigner de ce type de parole, surviennent les challenges du point de vue recherche. Le premier paragraphe de ce chapitre rappelle quelques chiffres sur les performances des systèmes en fonction de quelques tâches de lecture qui restent effectuées par des personnes sans troubles vocaux. Vient ensuite un paragraphe qui traite d'une étude sur l'intelligibilité des personnes saines et comment il peut être délicat de parler d'une intelligibilité parfaite pour ces personnes. Ensuite, je parlerai de deux travaux qui permettent de manipuler des données même « imparfaites » pourtant issues de parole « normale » pour améliorer les systèmes de traitement automatique. Enfin, je parlerai des travaux que j'ai co-encadré autour d'un corpus cette fois-ci enregistré par des personnes saines, mais non natives du français.

2.2 Contours de la parole « standard »

Dans toute recherche scientifique, l'étalonnage est une phase importante. Le scientifique doit s'interroger sur ce qu'est sa référence et en parole, nous devons pouvoir définir les contours d'une parole « canonique » , « normale » , « saine » ou encore « typique » . Dans la suite du chapitre, on s'autorisera à enlever les guillemets autour de ces notions pour ne pas alourdir le propos, même si ces notions sont sujettes à interprétation. Les premiers programmes d'enregistrements de grands corpus, notamment en français, prévoient d'enregistrer des centaines de personnes sur des tâches de lecture afin de garantir une bonne couverture phonémique et sur de la parole considérée comme la plus standard et ne mettant pas en défaut les premières modélisations, afin de s'attarder dans un premier temps au premier défi du taux de reconnaissance atteignable sur une parole la plus épurée possible. Des corpus tels que BREF [Lamel 1993] sont basés sur la lecture du journal Le Monde afin de pouvoir disposer d'une version transcrite assez aisément et être utilisés quasi directement par les modèles de SRAP. On utilise alors le terme de parole « propre » pour décrire ce type de production. Ces corpus qui étaient les premiers disponibles ont constitué le socle de la parole standard, du point de vue du TAP. S'écarter de la parole lue par une personne de sexe masculin est devenu au fil des avancées technologiques, un défi scientifique à relever. Sans parler des corpus acquis à des fins de Dialogue Oral Homme-Machine où la phraséologie est assez limitée, les corpus suivants sont enregistrés sur une base semi-préparée comme le corpus ESTER [Gravier 2004], issu de parole radiophonique et journalistique.

Cette parole, produite pour la plupart par des journalistes masculins, professionnels de la production orale, est pourtant déjà sujette à de la variabilité. En effet, ce type de discours est teinté de différents degrés de spontanéité. Dans [Jousse 2008], les auteurs ont segmenté et annoté 11h de parole provenant de du corpus ESTER en 10 degrés de spontanéité, allant de la parole lue totalement préparée et « propre » à de la parole très spontanée comportant de nombreuses disfluences. Sur l'ensemble des 11h, le système du LIUM [Deléglise 2005] obtient 25,1% d'erreur mot. Le tableau 2.1 montre que déjà sur ce type de parole très proche en terme de production et propre, le SRAP réagit de manière différente.

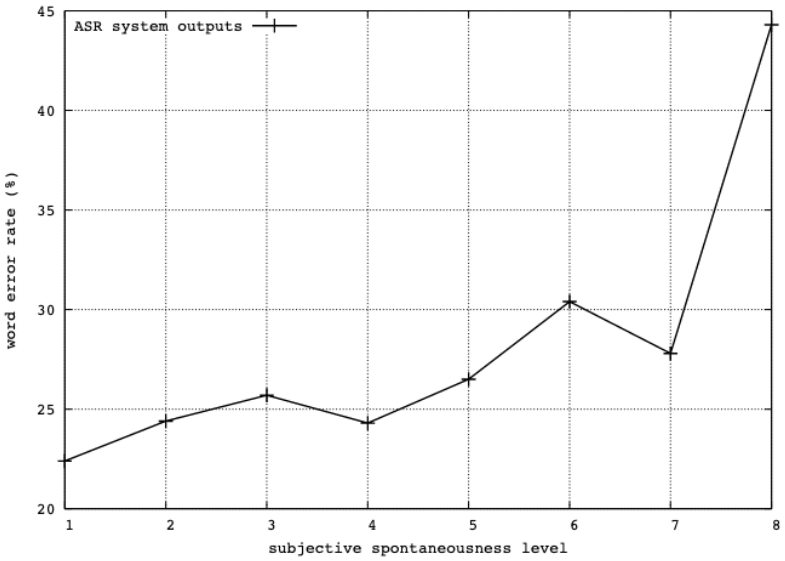

FIGURE 2.1 – Taux d'erreur mot en fonction du degré de spontanéité, tableau issu de [Jousse 2008]

À ces degrés de spontanéité causés par les disfluences vont s'ajouter les artéfacts liés à l'émotion ressentie par le locuteur, l'accent régional... Toutefois, ce type de production, avec ou sans artéfact est encore un type de production possible et admis comme standard dans une langue donnée. A quel moment franchissons-nous le seuil de l'« atypisme » ? À que moment s'écarte-t-on d'une parole saine ou normale ?

Par exemple, en apprentissage de langue seconde, la parole normale est de la parole d'un locuteur natif. L'apprenant, tout en n'ayant pas de problème de communication, va tout de même produire une parole plus déviante, et considérée jusqu'à un certain point comme perfectible. Dans ce cas, le système de TAP va chercher à analyser la déviance et proposer en fonction de la tâche, différents scénarios. Pour la transcription de la parole, le locuteur standard (celui pour lequel le système va avoir les meilleures performances) est un homme d'environ 35 ans et une personne plus âgée ou plus jeune va obtenir de plus faibles taux de reconnaissance par le système. Leur parole n'étant pour autant pas déviante, mais sortant uniquement du cadre de l'apprentissage du modèle canonique. En contexte de pathologie, la parole étalon est celle d'un locuteur « sain » , dans ce sens où celui qui produit l'énoncé n'a pas la pathologie vocale étudiée.

En s'attardant sur le contexte de la pathologie, on peut tout de même questionner ce concept de parole saine, ou parole étalon. Le paragraphe suivant propose un retour sur les travaux de recherche de Timothy Pommée [Pommée 2021c] que j'ai co-encadré et qui s'est interrogé sur l'intelligibilité d'une parole saine au détour d'une revue systématique de la littérature dans le contexte de la pathologie.

2.3 Intelligibilité d'une parole saine

Malgré son caractère hautement variable, la parole saine est considérée le plus souvent comme parfaite avec une intelligibilité de 100%. Dans l'article [Van Lierde 2012], les auteurs écrivent d'ailleurs : « Le groupe contrôle avait une intelligibilité normale » . Pourtant, de nombreuses études s'accordent sur la

variabilité de la parole saine, pouvant être affectée par le stress, l'émotion, la fatigue, l'accent régional ou encore la coarticulation, phénomènes qui amoindrissent l'intelligibilité du locuteur [Benzeghiba 2007, McCloy 2015]. Ceux-ci entraînent un écart par rapport à une parole canonique, car ils en altèrent soit la vitesse, soit la précision (« speed-accuracy trade-off ») [Guenther 1995, Meunier 2007, Tremblay 2017]. Lindblom explique par la théorie de l'« hyper/hypo-parole » [Bond 1994, Lindblom 1990] cette dégradation d'une intelligibilité maximale par différentes stratégies acoustiques et phonétiques. En comprenant ces stratégies, nous pourrons mieux appréhender ce qui relève des contraintes de la parole spontanée en contexte de communication naturelle, et quelles déviations indiquent elles une parole altérée. Dans sa revue systématique de la littérature, Timothy Pommée indique que plusieurs études, notamment sur le vieillissement physiologique [Hazan 2017, Kuruvilla-Dugdale 2020], montrent qu'une grande partie de la variabilité de la parole saine est imputable à des caractéristiques acoustiques-phonétiques spécifiques du locuteur [Bradlow 1996, Metz 1990]. Notre travail dans cette revue est consacré aux mesures acoustiques employées dans l'investigation de l'intelligibilité en parole saine afin de mieux anticiper et comprendre les comportements des mesures une fois appliquées à la parole pathologique. Les données de cette revue confirment la nature variable et « imprécise » de la parole chez les locuteurs adultes sains. L'observation principale est que parmi les études utilisant le pourcentage d'identification correcte, quatre articles ont trouvé des valeurs supérieures à 90% (sur des mots, des voyelles isolées et des voyelles dans des syllabes consonne-voyelle-consonne [CVC]), quatre autres ont trouvé des scores moyens entre 60,6% et 71% (sur des phonèmes dans des syllabes CVC et sur des syllabes). La variabilité de la parole chez les locuteurs sains est également observée entre les sujets à l'intérieur des différentes études. Par exemple, alors que trois des études utilisant des pourcentages d'identification correcte rapportent un écart-type relativement faible (variant de 1,12% à 4%), les études utilisant des échelles ordinales montrent une plus grande variabilité : si tous les résultats sont normalisés en pourcentages, les écarts-types varient de 6,25% à 12%. Bien qu'une part de cette plus grande variabilité puisse être liée au type d'échelle employée, ces résultats tendent tout de même à illustrer que même chez les locuteurs sains, les limites physiologiques ne permettent pas toujours au système de production de la parole de répondre aux nombreuses exigences de la parole spontanée. Les « imprécisions » qui en résultent se situent principalement au niveau des phonèmes [Rossi 1998, Schiller 2006], entraînant un certain recouvrement des catégories phonémiques, c.-à-d. des réductions vocaliques et consonantiques, ainsi que des omissions de phonèmes [Benzeghiba 2007, Guenther 1995, Meunier 2007, van Son 1996, van Son 1999].

Ce constat, selon lequel la parole des personnes saines ne garantit pas une intelligibilité parfaite, permet une utilisation spécifique des outils de technologie vocale. En se concentrant sur la parole saine et en observant les différences par rapport à un ensemble de paroles canoniques ou plutôt à l'ensemble des enregistrements de parole utilisés pour entraîner les modèles acoustiques, plusieurs systèmes peuvent être déployés en fonction des besoins. Dans la section suivante, un système de reconnaissance de la parole, entraîné sur des données de parole saine, utilise ses modèles internes (les modèles acoustiques et le modèle de langage) pour générer un auto-diagnostic de ses propres sorties : les mesures de confiance introduites dans le chapitre précédent. Celles-ci sont ici utilisées afin d'obtenir un corpus d'entraînement plus large avec plus d'exemples en filtrant un corpus non annoté, en le transcrivant automatiquement et en prenant les exemples dont la transcription avait un fort taux de mots supposés corrects par les mesures de confiance. Ces exemples sont réinjectés dans un nouveau corpus d'apprentissage.

2.4 Applications de mesures de confiance

2.4.1 Pour le filtrage de séquences de mots corrects

Nous avons vu que même sur la parole saine, plusieurs différences de niveaux d'intelligibilité même subtiles existent. En prenant en compte ce phénomène, on peut utiliser les mesures de confiance vues au chapitre précédent (1.4.2)pour mesurer et mettre à profit ces différents niveaux d'intelligibilité. Dans cette section, nous prenons comme point de départ un système de reconnaissance automatique de la parole entraîné sur de la parole saine. La parole canonique (ayant une intelligibilité étalon) est donc dans notre exemple la parole d'un premier corpus d'apprentissage pour lequel nous disposons de transcriptions humaines. Un corpus non annoté est également disponible. Les mesures de confiance calculées grâce à l'auto-diagnostic de notre système sur les sorties obtenues pour ce corpus peuvent servir de mesures de distance à la « norme » donnée par le premier corpus. À partir des sorties obtenues et des mesures de confiance associées, les sorties présentant un taux élevé de mots supposés corrects (aka dont la combinaison des mesures de confiances est élevée) peuvent être sélectionnées et intégrées dans un nouveau corpus d'apprentissage comme décrit dans la figure 2.2.

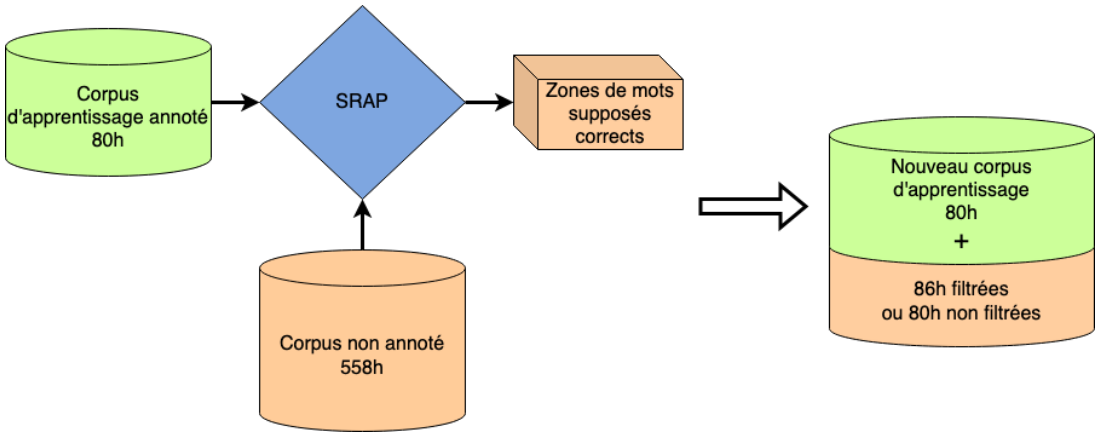

FIGURE 2.2 – Principe du filtrage de séquences de mots corrects : une première passe de reconnaissance automatique de la parole est effectuée sur un corpus non annoté, résultant en une labelisation de mots comme étant possiblement corrects ou non. Ces mots sont ensuite injectés dans le premier corpus d'apprentissage pour l'augmenter

Des expériences ont été réalisées sur les corpus de la campagne d'évaluation ESTER [Gravier 2004] afin de voir si le fait de filtrer le corpus non annoté en ne prenant que les zones de mots corrects résulte en de meilleures performances finales par rapport à ne pas filtrer ce corpus et l'ajouter au premier corpus comme décrit dans [Mauclair 2006b]. Une passe de reconnaissance est effectuée sur un corpus et les mesures de confiance permettent d'en sélectionner les séquences de mots dont l'annotation est plausiblement correcte. Ces séquences de mots constituent alors un corpus additionnel qui vient augmenter le corpus d'apprentissage de manière non supervisée, corpus additionnel qui comporte le minimum de mots incorrects pouvant « bruiter » le corpus d'apprentissage. Ce corpus additionnel provient d'un corpus fourni par la campagne ESTER non transcrit dont 558h ont été transcrites automatiquement par le SRAP utilisé lors de ces expériences.

Dans le tableau 2.1, nous pouvons constater que l'ajout aveugle de 80h améliore de 0,2% le taux

TABLEAU 2.1 – Taux d'erreur obtenu sur le même corpus de test avec l'adjonction d'un corpus filtré ou non filtré à un premier corpus pour l'apprentissage des modèles acoustiques

Corpus d'apprentissage	**Taux d'erreur (%) sur le Test d'ESTER**
Initial 80h (Ω)	23,7%
Ω + 80h non filtrées	23,5%
Ω + 86h filtrées	**23,2**%

d'erreur mot alors que l'ajout de 86h filtrées à l'aide d'un seuil sur les mesures de confiance permet une amélioration significative de 0,5%.

Ne conserver que des zones de paroles dont l'intelligibilité est proche de celle du corpus de départ permet ici d'améliorer les performances d'un système de SRAP.

Dans la partie suivante, nous verrons une deuxième application de ces mesures de confiance, toujours sur de la parole saine, pour permettre la combinaison de systèmes de reconnaissance de la parole.

2.4.2 Pour la combinaison de systèmes de reconnaissance

Dans les travaux effectués avec [Lecouteux 2007], j'ai utilisé la mesure de confiance finale précédente pour faire de la combinaison de systèmes de reconnaissance. La combinaison de systèmes proposée est basée sur une méthode par décodage guidé. Le décodage guidé consiste à effectuer une première passe de reconnaissance automatique de la parole, en utilisant un système auxiliaire (ici le système du LIUM [Deléglise 2005]) qui propose ses meilleures hypothèses. À chaque mot de l'hypothèse est associé un score de confiance (celui de [Mauclair 2006a]). Cette information permet ensuite de réévaluer dynamiquement les probabilités du modèle de langage au sein de l'algorithme de recherche du système primaire (le décodeur SPEERAL du LIA [Nocera 2002]). La combinaison des systèmes est illustrée sur la figure 2.3.

Dans ces travaux, nous indiquons que cette combinaison utilisant une ré-estimation dynamique des probabilités linguistiques à laquelle nous ajoutons une adaptation croisée des modèles acoustiques permet d'obtenir un gain en WER absolu de 1,9% par rapport aux deux systèmes de référence sur le corpus ESTER. De plus, l'analyse des résultats montre que la combinaison génère de nouvelles hypothèses correctes qui ne sont présentes dans aucun des systèmes de référence.

En conclusion, la parole saine est donc la parole cible sur laquelle les modèles des systèmes de traitement automatique de la parole sont appris. Il s'agit de parole « propre » , enregistrée dans des conditions sans bruits environnementaux et avec du matériel difficilement transposable en situation « réelle » . Comme nous l'avons vu, la parole propre est déjà suffisamment variée pour fournir un cas d'étude intéressant et riche pour améliorer des systèmes de traitement automatique de la parole. Dans le chapitre précédent, nous avons noté que des altérations peuvent subvenir dans le cadre d'une mise en situation réelle. Il peut s'agir du canal de transmission qui provoque une dégradation, mais également, et c'est ce qui nous intéresse ici, d'une production de parole qui dévie d'une parole canonique. Que se passe-t-il alors quand les systèmes de traitement automatique de la parole sont confrontés à de la parole « saine » , c'est-à-dire ou le locuteur n'a pas de pathologie vocale, mais où cette parole s'écarte de la parole « propre » et « typique » d'une langue ?

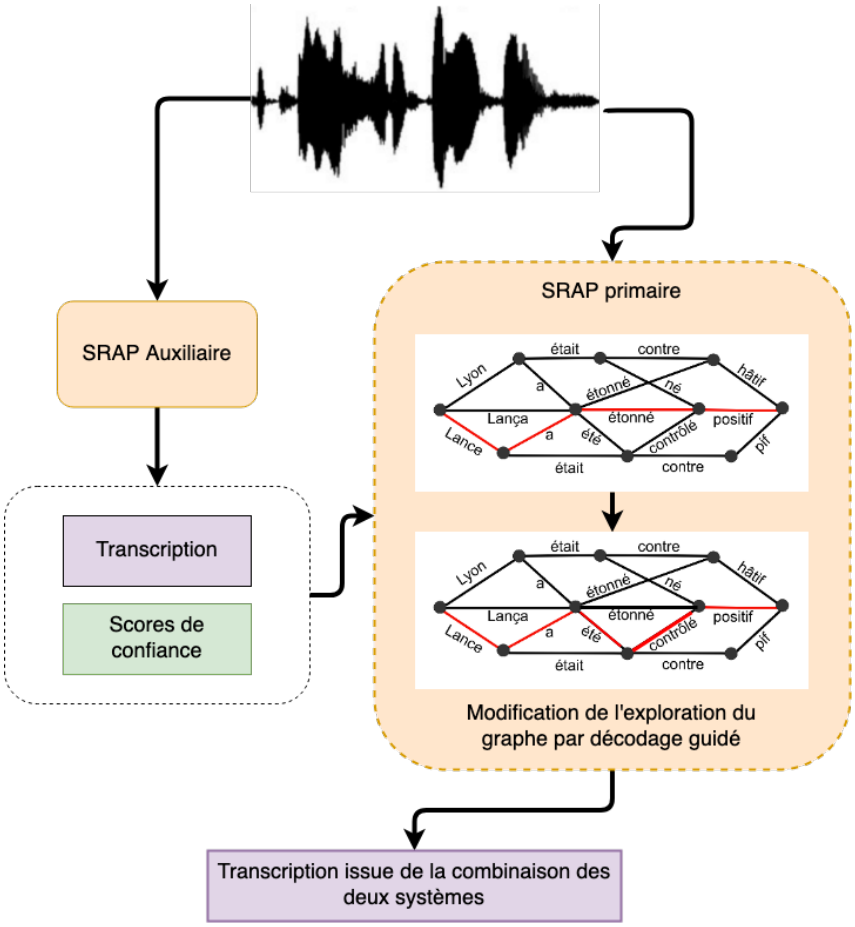

FIGURE 2.3 – Principe de la combinaison par décodage guidé : La combinaison proposée consiste à effectuer une première passe de reconnaissance automatique de la parole, en utilisant un système auxiliaire qui propose ses meilleures hypothèses ainsi qu'un score de confiance sur celles-ci

2.5 Quand la réalité s'écarte de la « norme »

Comme nous l'avons vu, le caractère normal d'une parole désigne plutôt, du point de vue du traitement automatique de la parole, le fait que celle-ci corresponde à la parole sur laquelle ont été appris les modèles utilisés. D'un corpus considéré « simple » au départ, sans artéfacts, sans bruits, provenant de locuteurs qui maîtrisent leur discours dans leur langue maternelle, les chercheurs ont pu obtenir des systèmes aujourd'hui plus que performants. En s'éloignant petit à petit de cette parole standard, les chercheurs ont commencé à s'intéresser à une parole différente, de plus en plus éloignée du type de parole considéré en premier lieu.

Que fait-on de ce nouveau type de parole ? Réapprend-on les modèles avec des corpus différents pour avoir des taux d'erreur de plus en plus réduits ou se sert-on du modèle de parole standard pour mesurer la différence entre la nouvelle parole et la standard ? Cela dépend des applications.

Par exemple, dans les travaux de [Laborde 2016], nous nous posons la question de l'évaluation de la prononciation du français par des apprenants japonais. Ici, la norme n'est pas tant la façon dont le locuteur s'exprime dans sa langue maternelle, mais plutôt comment sa prononciation est dégradée dans

la tentative de production d'un son dans une langue seconde et donc différente d'une production d'un locuteur natif. L'algorithme retenu pour évaluer cette déviance de prononciation est basé sur le Goodness Of Pronunciation [Witt 2000]. Comme illustré dans la figure 2.4, cet algorithme évalue la différence entre les sorties de deux systèmes de décodage de la parole. Le premier système est un système contraint par le texte où l'alignement fourni est une segmentation temporelle des phones attendus par rapport au signal audio. Le deuxième système est une reconnaissance libre qui détermine la séquence de phone la plus probable correspondant à l'extrait audio.

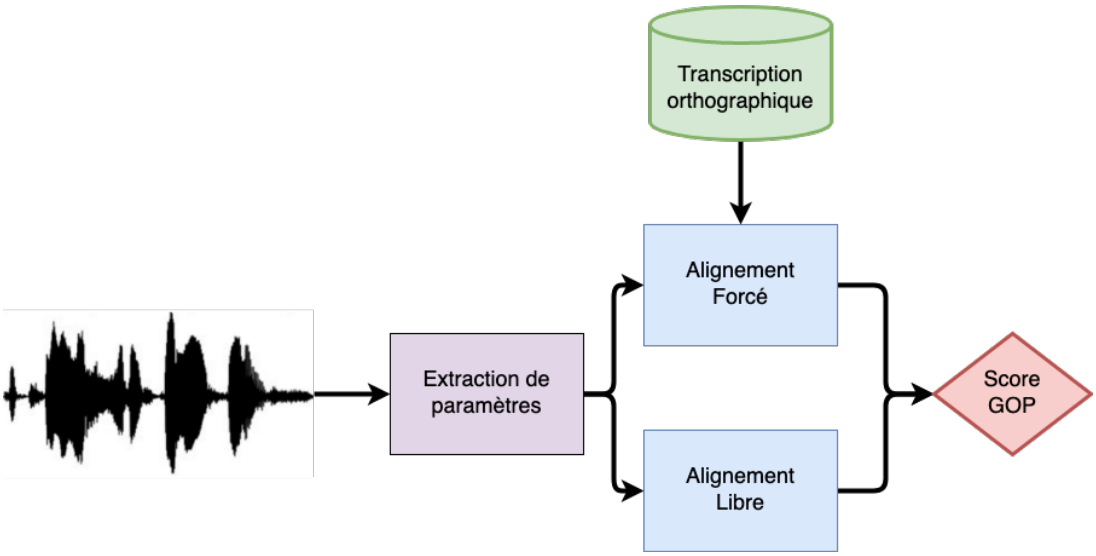

FIGURE 2.4 – Principe de l'algorithme du GOP avec les phases d'alignement forcé et d'alignement libre adapté de [Witt 2000]

Pour chaque phone p aligné sur O^p observations d'une durée totale de $\mathrm{NF}(p)$ trames, un score GOP est calculé suivant l'équation 2.1. Il s'agit de la valeur absolue du rapport entre le log vraisemblance de la phase d'alignement forcé $p(O^p|p)$ et celui de la phase d'alignement libre $\max_{\text{path}} p(O^p|\text{path})$. Le GOP est normalisé par la durée en termes de nombres de trames $\mathrm{NF}(p)$ d'un segment d'un phone obtenu par l'alignement forcé (segment de référence). Quand un seul phone de l'alignement libre apparaît au cours d'un segment de référence et qu'il est le même que celui de l'alignement forcé, le score GOP est de zéro. Dans le cas contraire, plus le score est élevé, plus il est probable que le phone ait été mal prononcé.

$$GOP(p) = |\log(\frac{p(O^p|p)}{\max_{\text{path}} p(O^p|\text{path})}) * \frac{1}{\mathrm{NF}(p)}| \tag{2.1}$$

L'expérience relatée dans [Laborde 2016] a eu lieu dans le cadre du projet PHON-IM. Celui-ci a pour objectif d'étudier l'évolution des compétences en perception et production de la parole chez des apprenants japonais débutants venant faire un stage d'immersion linguistique en France. 24 apprenants ont ainsi été enregistrés à leur arrivée en France, puis un mois après, après avoir suivi des cours de français et participé à des ateliers de correction phonétique (entraînement à la perception et à la prononciation des sons du français). 71 pseudo-mots et 9 phrases ont été répétés par chacun d'eux à chaque séance pour plus de 58 minutes d'enregistrement.

Afin de repérer automatiquement cette mauvaise prononciation, il faut réussir à définir un seuil du score GOP au-dessus ou en dessous duquel la prononciation sera catégorisée comme correcte ou incorrecte. Pour ce faire, il faudrait idéalement disposer d'un corpus de parole non native annoté manuellement au niveau phonémique. Cependant, la taille de ces ensembles de données est généralement moindre que celle d'un corpus de discours natif utilisé pour apprendre les modèles acoustiques d'un SRAP. Ainsi, une technique adoptée dans la littérature consiste à introduire des erreurs de prononciation artificielles en substituant des transcriptions de phones dans le lexique de prononciation utilisé lors du calcul du score GOP [Witt 2000, Kanters 2009]. Nous avons également utilisé cette méthode pour tirer parti d'un vaste corpus de locuteurs natifs français, le corpus BREF [Lamel 1993]. Comme nos locuteurs cibles sont des locuteurs natifs japonais apprenant le français langue étrangère (FLE), nous nous sommes concentrés sur les deux phonèmes français /R/ et /v/, qui ont été signalés comme étant problématiques pour les locuteurs japonais [Tomimoto 2008]. Les confusions les plus fréquentes se produisent entre /R/ et /l/ [Yamasaki 1999] , et entre /b/ et /v/ [Detey 2005]. Ainsi, chaque /l/ du lexique de prononciation a été remplacé par /R/ (de sorte que l'ASR s'attende à un son [R] et obtienne un [l] dans l'audio), et de la même manière, chaque /b/ a été remplacé par un /v/. Pour chaque phone cible, un seuil a été calculé en recherchant le score qui minimise le nombre d'erreurs, c'est-à-dire la somme des fausses acceptations et des faux rejets.

Le corpus de test est quant à lui composé de parole lue recueillie auprès d'un groupe homogène d'étudiants japonais en FLE, le corpus PHON-IM. Un total de 414 /R/ et 368 /v/ ont été étiquetés comme correctement ou incorrectement prononcés par deux annotateurs ayant une solide formation en phonétique et une expérience dans la transcription de la parole dans le contexte de l'enseignement du FLE (voir le tableau 2.2). Un taux d'inter-annotation élevé de 84,4% a montré un large consensus dans leur annotation, avec un accord plus important sur les réalisations /v/ que sur les réalisations /R/ : 86,1% et 82,9%, respectivement.

Tableau 2.2 – Nombre d'occurrences de /R/et de /v/ dans le corpus PHON-IM annotés de façon similaire par les deux phonéticiens

Phone	PHON-IM	
	#correct	#incorrect
/R/	215	128
/v/	267	50

L'objectif de ce travail était d'améliorer l'approche GOP. Pour ce faire, nous avons utilisé le forced-aligned GOP, appelé F-GOP [Pellegrini 2015] et implémenté avec HTK [Young 1994]. La différence avec le GOP originel tient dans la prise en compte des frontières du phone de la phase d'alignement forcé pour contraindre les frontières de la phase de reconnaissance libre. Nous avons également ajouté des informations aux scores F-GOP sous la forme de caractéristiques supplémentaires données comme entrées à un modèle probabiliste, un modèle de régression logistique (LR) :

1. l'identité du phone attendu ;
2. l'identité du phone reconnu ;
3. Les scores de vraisemblance obtenus lors de la phase d'alignement forcé et lors de la phase de reconnaissance libre ;

4. Le nombre de traits distinctifs différents entre les phones trouvés dans chaque phase du calcul du score GOP ;
5. l'identité des phones gauche et droit ;
6. Le ratio entre la durée du phone et la durée de l'état central du HMM qui est supposé le plus stable et le plus long.

Nous avons entraîné les classificateurs LR sur les mêmes corpus sur lesquels les seuils ont été fixés pour la méthode de base (BREF avec erreurs artificielles). Les poids du modèle LR fournissent des informations sur l'importance relative des caractéristiques d'entrée. Le poids estimé de la caractéristique du score GOP était de -0,633, une valeur négative qui correspond au fait que plus le score GOP est élevé, plus une erreur de prononciation est probable. Les poids pour l'identité catégorielle du phone étaient de 0,627 et 0,445 pour /v/ et /R/, respectivement. Le poids de /v/ est légèrement supérieur à celui de /R/, ce qui est également cohérent avec le fait que le seuil GOP correspondant est plus élevé pour ce phone (1,13 et 2,97 pour /R/ et /v/, respectivement).

Le tableau 2.3 montre les résultats obtenus avec les approches de base F-GOP, ainsi qu'avec les différents modèles LR, en utilisant uniquement les scores F-GOP et l'identité phone attendu (deuxième colonne), et en ajoutant chacune des cinq caractéristiques supplémentaires, une à la fois. La dernière colonne donne les résultats de la meilleure combinaison de caractéristiques. Dans chaque cellule du tableau, deux nombres sont donnés pour /R/ et /v/, respectivement. L'approche baseline F-GOP a donné une précision de 63,8 %. Le modèle LR avec ajout de l'identité du phone attendu a donné une performance similaire de 64,4 %. En analysant les résultats pour /R/ et /v/ séparément, il apparaît que lorsque le phone reconnu correspond à celui attendu, les deux systèmes prédisent toujours les prononciations comme correctes. Cinquante-cinq pour cent des 343 réalisations attendues de /R/ ont été reconnues comme [R], et les substitutions les plus fréquentes impliquaient [f] (13 %) et le modèle pour les pauses (9 %). Ceci est cohérent avec les annotations manuelles, qui ont montré que les réalisations de /R/ étaient le plus souvent transcrites en utilisant le téléphone japonais [h] - une consonne non voisée, grave et fricative plutôt proche de [f] ou d'une pause respiratoire. Pour /v/, 25% et 41% des occurrences ont été reconnues comme [v] et [f], respectivement. Seulement 1% des occurrences ont été reconnues comme [b], ce qui est en contradiction avec les données manuelles : [b] était le phone alternatif le plus fréquent utilisé par les annotateurs pour transcrire les productions des locuteurs japonais.

TABLEAU 2.3 – Scores (en %) du classifieur sur le corpus PHON-IM en utilisant les différents paramètres sur les phones /R/ et /v/.

	Baseline	Régression logistique						
	F-GOP	+1	+2	+3	+4	+5	+6	2+4+5
SA (%)	68,5/58,7	71,1/57,1	68,5/81,4	69,1/54,9	69,7/63,7	73,2/57,1	70,8/57,4	69,1/85,8
Global(%)	63,8	64,4	74,4	62,4	66,8	65,6	64,5	76,9

2.6 Conclusion

Dans ce chapitre, nous avons vu que finalement dans le domaine de l'altération de la production, la parole est un continuum qui va de la parole standard à la parole fortement altérée en passant par la parole atypique, c'est-à-dire celle à laquelle ne s'attend pas notre système de traitement automatique. Même en

parole saine, nous avons vu que la norme est quelque chose de relatif et que les personnes dites contrôles dans les études n'ont pas une référence égale et optimale. L'intelligibilité a été dans ce chapitre un moyen de s'intéresser à la prononciation parfaite pour montrer les précautions qu'un chercheur doit prendre en commençant une étude et en décrivant sa population contrôle. Les systèmes eux-mêmes peuvent gagner beaucoup en généralisation pour par exemple faire croître de manière non supervisée leur corpus d'apprentissage avec des corpus nouveaux. Les systèmes peuvent aussi utiliser des modèles spécifiques à chaque altération en prenant en compte des indices ou des scores dans la chaîne de traitement. À mesure que l'on s'écarte d'une prononciation standard, on peut utiliser des méthodes de TAP pour exprimer cette déviation et, comme avec l'exemple de la L2, concevoir des systèmes qui indiquent à quel point une prononciation s'écarte d'une parole typique. Et si l'objectif n'est pas de transcrire la parole atypique, on peut vouloir implémenter un logiciel d'aide à la prononciation, d'aide à la lecture pour les jeunes enfants... À l'autre extrême du continuum, il y a la parole pathologique. Dans le prochain chapitre, je montrerai les différents projets qui m'ont amené à étudier ce type de parole et ce qui en a résulté aussi bien pour de l'aide au diagnostic que pour la compréhension générale du déficit de communication.

CHAPITRE 3

QUID DE LA PAROLE PATHOLOGIQUE ?

SÉLECTION DE PUBLICATIONS RELATIVES À CE CHAPITRE :

- Woisard, V. and Astesano, C. and Balaguer, M. and Farinas, J. and Fredouille, C. Gaillard, P. and Ghio, A. and Giusti, L. and Laaridh, I. and Lalain, M. and Lepage, B. and **Mauclair, J.** et al. *C2SI corpus : a Database of Speech Disorder Productions to Assess Intelligibility and Quality of Life in Head and Neck Cancers* Dans : Language Resources and Evaluation, Springer, janvier 2020

- Pommée, T. and Balaguer, M. and **Mauclair, J.** and Pinquier, J. and Woisard, V. *Assessment of adult speech disorders : current situation and needs in French-speaking clinical practice* Dans : Logopedics Phoniatrics Vocology, pp1-15, janvier 2021

- Pommée, T. and Bouvier, L. and Pinquier, J. and **Mauclair, J.** et al. *Le voyage d'Alice : un texte standardisé pour l'évaluation de la parole et de la voix en Français*, Dans : Glossa, pp6-43, janvier 2024

- Quintas, S. and Abad, A. and **Mauclair, J.** and Woisard, V. and Pinquier, J. *Towards Reducing Patient Effort for the Automatic Prediction of Speech Intelligibility in Head and Neck Cancers*, Dans : ICASSP 2023

- Quintas, S. and **Mauclair, J.** and Woisard, V. and Pinquier, J. *Automatic Prediction of Speech Intelligibility based on X-vectors in the context of Head and Neck Cancer*, Dans : Interspeech 2020

Sommaire

3.1 Introduction

Plusieurs projets m'ont permis de travailler sur le domaine de la parole pathologique.

Lors de ma nomination en tant que Maître de Conférences sur l'Université Paris Descartes, j'ai travaillé à la création d'un corpus de paralysés faciaux avec Mme Peggy Gatignol, orthophoniste à l'hôpital de la Pitié Salpétrière. Une grosse partie de ce projet a consisté en l'enregistrement en salle d'audiométrie de l'hôpital d'un corpus de parole de patients paralysés faciaux (voir paragraphe 3.2.1).

Plus tard, j'ai contribué au montage du projet ANR Voice4PD-MSA dont l'objectif est double, à savoir développer un marqueur numérique objectif non invasif pour aider au diagnostic différentiel précoce entre la maladie de Parkinson et la maladie de l'atrophie multi-systématisée à syndrome Parkinsonien, et au diagnostic de Parkinson par rapport aux témoins sains (voir paragraphe 3.2.3).

Le projet INCA C2SI [1] puis le projet ANR RUGBI [2] qui a suivi m'ont permis de participer à l'élaboration d'un index automatique de sévérité/intelligibilité de voix de patients ayant un cancer des voies ORL puis de me focaliser sur la mesure objective du déficit d'intelligibilité (voir paragraphe 3.2.2). Le projet RUGBI propose notamment de développer un nouvel outil d'évaluation objective de l'intelligibilité basé sur i) l'identification des unités linguistiques pertinentes d'un point de vue acoustique et prosodique, et ii) l'identification de tâches linguistiques sensibles. Les partenaires sont le CHU de Toulouse, le laboratoire LNPL de l'Université Jean Jaurès de Toulouse, le laboratoire Parole et Langage de l'université d'Aix-Marseille et le laboratoire d'Informatique d'Avignon. Les chercheurs participant à ces projets sont des phonéticiens, des linguistes, des informaticiens, des orthophonistes et des médecins.

1. http://petra.univ-tlse2.fr/spip.php?article248
2. https://www.irit.fr/rugbi/

J'ai ensuite participé au montage du projet européen TAPAS [3] : Training Network on Automatic Processing of PAthological Speech. Mon implication au niveau administratif est décrite plus en détail dans la partie CV du document. En ce qui concerne la partie scientifique, il s'agit d'un réseau pluridisciplinaire Horizon 2020 financé par le programme Actions Marie Skłodowska-Curie de l'Union européenne. Ce projet a permis de former et d'encadrer 15 jeunes chercheurs/doctorants autour de collaborations interdisciplinaires et internationales entre partenaires académiques, industriels, associatifs et hospitaliers sur le domaine du traitement automatique des troubles de la parole. Ce projet européen a offert de nombreuses opportunités de formation, notamment par les séminaires (« training events ») organisés tous les 6 mois par l'un de ses partenaires. Ceux-ci visaient non seulement à développer les connaissances théoriques et appliquées des jeunes chercheurs (p. ex. en s'assurant que chacun d'eux a une connaissance de base des troubles de la parole et de la prise en charge orthophonique, ainsi que de l'informatique et de l'éthique de la recherche), mais aussi à les préparer pour l'après-thèse (p. ex. grâce à un séminaire portant sur l'entrepreneuriat). Enfin, le réseau TAPAS a fortement encouragé la participation de ses membres à des conférences internationales pour la diffusion des savoirs et le développement de réseaux professionnels. L'objectif général était l'amélioration de la qualité de vie des patients atteints de troubles de la parole, et plus spécifiquement, le projet se déclinait en trois volets : (a) la « détection », visant le développement de nouvelles techniques de traitement du signal audio pour la détection précoce et non invasive des troubles de la parole ; (b) la « thérapie », dont l'objectif était le développement de nouveaux outils de prise en charge automatisés ; (c) l'« assistance à l'autonomie à domicile », visant à améliorer les outils basés sur le traitement du signal audio pour améliorer l'autonomie des patients atteints de troubles de la parole. C'était dans le deuxième volet de ce programme international, celui de la « thérapie », que se sont inscrites les deux thèses avec lesquelles j'ai commencé l'encadrement de doctorants.

Tous ces projets tournaient donc autour de l'étude de la parole pathologique que je vais aborder par la suite en 4 paragraphes se référant à des axes de mon travail.

La première partie décrit les différents corpus sur lesquels j'ai travaillé en s'attardant plus spécialement que la phase de constitution et dans quel contexte ils ont été pensés. La deuxième partie traite des outils d'évaluation de la voix disponibles en milieu clinique et qui vont être utiles pour l'annotation des corpus de voix et de parole. Je traiterai ensuite de la caractérisation de la parole pathologique que j'ai pu étudier et des travaux qui ont été menés sur de la parole de patients aphasiques, des paralysés faciaux et des patients atteints de syndromes parkinsoniens. Enfin, je montrerai quelles modélisations ont pu être proposées dans plusieurs études sur ses différentes catégories de maladies, que ce soit pour de la classification ou de l'extraction de connaissances. L'extraction de connaissances offre, pour une maladie donnée la possibilité de savoir si un système automatique repère les troubles afférents à cette maladie et permet par exemple l'aide à la remédiation. La classification elle, permet de catégoriser un nouveau fichier audio parmi les maladies que l'on a à traiter et donc d'aider au diagnostic différentiel.

3.2 La constitution des corpus

Dans [Ghio 2021], les auteurs constatent la nécessité pour l'étude des troubles de la voix et de la parole de sortir du cadre de la recherche clinique. Étudier des pathologies induit l'observation d'un grand nombre de patients et de les confronter aux résultats établis sur de la parole « normale » . Cependant, obtenir des enregistrements seuls ne présente d'intérêt que s'ils sont combinés avec des caractéristiques cliniques

3. https://www.tapas-etn-eu.org/

du locuteur. L'objectif de [Ghio 2021] est alors de « proposer des recommandations pour la structuration de données sonores, physiologiques et cliniques dans le cas de corpus de parole issue de patients atteints de troubles de la voix et de la parole » . En suivant cette ligne directrice, je me suis retrouvée partie prenante de plusieurs constitutions de corpus qui avaient tous un objectif de recherche et pour lesquels il fallait réfléchir à quoi faire enregistrer au patient pour valider une hypothèse de recherche.

3.2.1 Paralysés faciaux

Le premier corpus de parole pathologique sur lequel j'ai travaillé et auquel j'ai contribué est celui de parole de paralysés faciaux. Ce corpus a été enregistré à l'hôpital La Pitié Salpétrière à Paris, France. La base de données contient 70 enregistrements sonores, des patients et des témoins sains. Les patients souffrent de paralysie faciale périphérique d'origine virale, due à des neurinomes (du nerf acoustique ou du nerf facial) ou encore d'origine traumatique. De plus amples détails sur les patients de ce corpus sont disponibles dans [Robert 2011]. L'échelle d'évaluation clinique la plus utilisée au niveau international est celle de House et Brackmann [House 1985] qui permet une cotation du degré de dysfonction faciale rapide et intégrant l'évaluation des séquelles, allant du grade I (fonctionnement normal) au grade VI (paralysie flasque totale). C'est une échelle à 6 niveaux qui permet uniquement de noter les aspects physiques de la paralysie (principalement le front, les yeux et la bouche) grâce à une observation directe effectuée par un praticien. Le tableau 3.1 met l'accent sur les difficultés au niveau de la bouche au travers des différents grades de l'échelle de House et Brackmann. Pour les patients atteints de paralysie faciale périphérique à partir du grade III, la littérature fait état d'un retentissement psychologique et de troubles fonctionnels tels que les troubles oculaires, les syncinésies, l'hémi-spasme, les troubles de la mastication, de la déglutition et de l'articulation [Gatignol 2004].
Certains des enregistrements proviennent du même patient au même stade ou à différents stades de la paralysie (voir tableau 3.1). Les patients ont été enregistrés dans une cabine insonorisée avec un microphone supercardioïde et sur un enregistreur numérique utilisant une résolution audio PCM WAV linéaire de 16 bits/44,1 kHz. Différents types de phrases ont été enregistrés. Les patients devaient lire différents textes : mots isolés, phrases isolées, textes de journaux, et ils devaient répondre spontanément à la question : « Comment fait-on une omelette ? ». Dans l'ensemble du protocole, la base de données consiste également en la mesure de la tension dans les lèvres des patients avec un dynamomètre, un test phonétique d'intelligibilité, un formulaire qui évalue la motricité des lèvres, de la langue et du visage et les patients doivent situer leur satisfaction articulatoire sur une échelle visuelle analogique (normalement utilisée pour la douleur, elle est ici utilisé pour la satisfaction).

3.2.2 Le corpus C2SI

Dans le cadre du projet Carcinologic Speech Severity Index (C2SI), une série d'enregistrements de la parole de patients souffrant de cancers des voies aéro-digestives supérieures a été collectée [Astésano 2018]. La visée de ce corpus est de pouvoir mesurer l'impact de cette pathologie sur la production de la parole [Woisard 2021] et d'évaluer la qualité de vie des patients après le traitement. Le corpus est composé d'enregistrements audio de 134 sessions avec les métadonnées associées telles que la taille et la localisation de la tumeur, le traitement administré... Plusieurs tâches ont été effectuées par les locuteurs comme la lecture de pseudo-mots [Ghio 2018], la lecture de phrases, une tâche d'évaluation des fonctions prosodiques [Nocaudie 2018], la lecture de texte ou encore une description d'image. Des évaluations per-

TABLEAU 3.1 – Les mouvements de la bouche référencés dans l'échelle de House et Brackman [House 1985]

Grade	**Caractéristiques**
I Normal	Fonction faciale normale
II Atteinte discrète	Légère paralysie. Au repos, symétrie et tonus normaux. **Bouche : légère asymétrie**
III Atteinte modérée	Asymérie évidente mais non choquante. Au repos, symétrie et tonus normaux **Bouche : faiblesse des muscles quand l'effort est maximum**
IV Atteinte modérément sévère	Faiblesse ou asymétrie défigurante Au repos, tonus et symétrie normaux **Bouche : asymétrique quand l'effort est maximum**
V Atteinte sévère	Mouvement à peine perceptible Asymétrie au repos **Bouche : mouvement très faible**
VI Paralysie totale	Aucun mouvement

ceptives de jurys naïfs et d'experts ont rendu disponibles trois indicateurs décrivant les caractéristiques de la parole des locuteurs enregistrés. Deux d'entre eux étaient basés sur l'évaluation subjective de six orthophonistes/phoniatres sur la tâche de description d'image : l'intelligibilité était définie comme la compréhensibilité d'un message basé sur le signal de parole, tandis que la sévérité était définie comme le degré de détérioration globale du signal (voir figure 3.1). Le troisième indicateur a été obtenu à partir d'un décodage acoustico-phonétique manuel des pseudo-mots [Lalain 2020]. Cet indicateur correspondait au nombre de caractéristiques qui différaient entre la forme attendue et l'élément transcrit en terme de classe acoustico-phonétique.

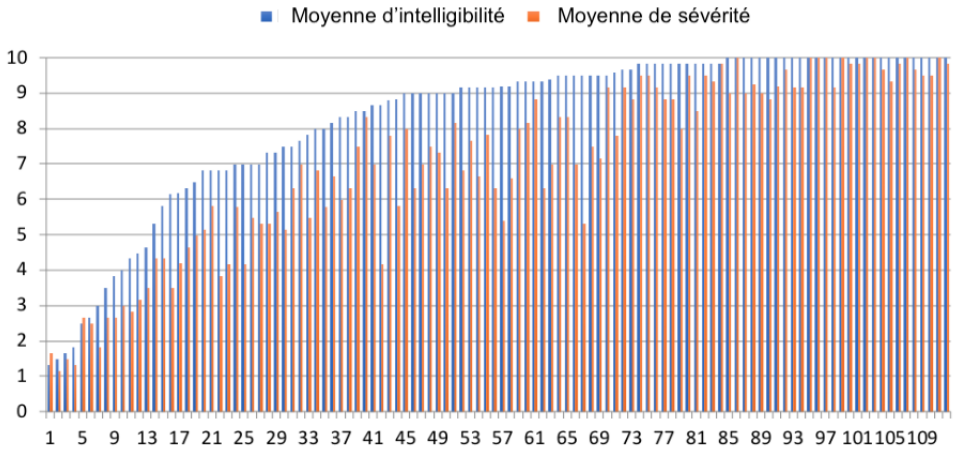

FIGURE 3.1 – Distribution des scores d'intelligibilité et de sévérité en ordonnée et numéro de sujets (par scores croissants d'intelligibilité) en abscisse

Les thèses de Timothy Pommée [Pommée 2021a] et Sebastião Quintas [Quintas 2022a] ont basé plu-

sieurs de leurs expériences sur ce corpus, notamment sur les tâches de lecture du passage de « La chèvre de Monsieur Seguin » d'Alphonse Daudet (LEC) et sur la tâche de lecture de pseudo-mots (DAP).

La tâche de lecture (LEC)

Pour la tâche LEC (voir encadré 3.2.1), le score d'intelligibilité est celui défini par la moyenne des scores que 6 experts orthophonistes/phoniatres ont fourni sur la tâche de description d'image, 0, correspondant à un message parfaitement intelligible et 10, correspondant à un message non intelligible. Le coefficient de corrélation intraclasse des juges est de 0.69.

Encadré 3.2.1 – Le passage de « La chèvre de Monsieur Seguin » d'Alphonse Daudet lu pour la tâche de lecture (LEC)

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 maître
ni la peur du loup rien ne les retenait.
C'était paraît-il des chèvres indépendantes
voulant à tout prix le grand air et la liberté.

La tâche de lecture de pseudo-mots (DAP)

Dans le contexte du corpus C2SI, tous les locuteurs ont enregistré 52 pseudo-mots différents qui respectent l'orthographe et la prononciation françaises [Ghio 2018]. L'ensemble de 52 pseudo-mots était différent pour chaque patient, suivant la structure suivante : $C(C)_1V_1C(C)_2V_2$, où $C(C)_i$ est une consonne isolée ou un groupe consonantique et V_i est une voyelle. L'encadré 3.2.2 présente quelques exemples des pseudo-mots utilisés.

Encadré 3.2.2 – Exemple d'un ensemble de 52 pseudo-mots

banfou bleja boucti brimpli chessant choniou
clifant cogu crimpin daillu dinrant dredi
fanrsi flinrpu fouma fravi gabi glunou grorvo
guchin joutu juro lanvin lerda messo mouco
nianlo niejo noksa nouillou pastu pidant
ploniou pripin psila quiga rinta rurnu
sanvrin scuna souquin spaclant sticho tangri
tougzu tradrou virjant vumou yainzi
yaltin zebou zouzant

La mesure perceptuelle d'intelligibilité utilisée pour la tâche DAP a été obtenue en moyennant le score de la transcription individuelle de chaque pseudo-mot par 3 auditeurs naïfs. Ce score perceptif est utilisé ici comme vérité terrain (référence). Il a été calculé en fonction de la distance entre le mot transcrit et le mot d'origine, en fonction d'une matrice de coût des voyelles et des consonnes [Ghio 2018]. Dans le cadre de ce travail, les scores perceptifs étaient compris entre 0, correspondant aux mots parfaitement intelligibles et 5, correspondant aux mots non-intelligibles. La moyenne des 3 auditeurs correspond au

score de chaque mot, et la moyenne des 52 mots de chaque patient correspond au score DAP de chaque patient respectivement.

Sauf autre mention, les expériences relatées dans la suite de ce chapitre sont effectuées sur un ensemble de 126 locuteurs (40 contrôles et 86 patients).

3.2.3 Voice4PD-MSA

Dans le cadre du projet Voice4PD-MSA, plusieurs enregistrements audio ont été effectués sur différentes tâches. Les patients de ce corpus souffrent de deux maladies neuro-dégénératives, Parkinson (MP) et l'atrophie multi-systématisée(AMS). En stade précoce, les symptômes de la MP et de l'AMS sont très similaires, en particulier chez les patients atteints d'AMS-P où le parkinsonisme prédomine. Par conséquent, le diagnostic différentiel entre AMS et MP est souvent très difficile à établir aux premiers stades de la maladie, alors que la certitude d'un diagnostic précoce est importante pour le patient en raison du pronostic divergent et de la gravité du pronostic de l'AMS. Aucun marqueur objectif validé n'est actuellement disponible pour guider le clinicien dans ce diagnostic et le besoin de tels marqueurs est donc très élevé dans la communauté neurologique. Comme la dysarthrie est un symptôme commun et précoce dans les deux maladies, mais d'origine différente, notre approche consiste à rechercher comment la caractériser, par le biais du traitement automatique de la parole et du signal, et rechercher des différences entre les patients atteints de MP et d'AMS aux premiers stades des maladies. En premier lieu, un diagnostic est posé par un neurologue pour confirmer dans quel groupe appartient le patient (MP ou AMS). Ensuite, l'orthophoniste conduit l'enregistrement dans une salle dédiée avec le même protocole dans les deux hôpitaux.

Les enregistrements du corpus obtenus lors de ce projet ont été réalisés dans les centres hospitaliers universitaires de Bordeaux et de Toulouse. À ce jour, 26 patients atteints de la maladie de Parkinson, 13 patients atteints d'AMS et 19 témoins sains sont disponibles dans ce corpus. Les sessions d'enregistrements ont été réalisées dans les salles de consultation à l'environnement sonore silencieux. Un enregistreur numérique de haute qualité (H4n) a été utilisé avec deux microphones, un microphone-casque et un microphone Rode NT1 cardioïde à environ 10 centimètres du locuteur et un microphone AKG C1000S cardioïde branché sur la station EVA (voir figure 3.2). Tous les microphones réalisent leurs enregistrements en parallèle afin d'éviter de répéter les mêmes tâches plusieurs fois pour chaque patient.

Durant la session d'enregistrement, tous les locuteurs ont réalisé les mêmes tâches de production de parole : lecture des premières phrases de texte « La chèvre de monsieur Seguin » , parole spontanée, voyelle /a/ tenue, diadococinésie (répétition de syllabes /pataka/ et /badaga/), lecture de logatomes (pseudo-mots). La durée totale d'une session d'enregistrement est d'environ 15 minutes par locuteur.

3.2.4 Le voyage d'Alice

Tous les corpus précédents sont destinés à permettre l'évaluation perceptive puis par des méthodes de traitement automatique de la parole des échantillons audio de différents niveaux de granularité.

Tant en pratique clinique qu'en recherche, les textes de référence permettent d'obtenir un aperçu rapide des caractéristiques de la parole du patient [Auzou 2006], avec moins d'hésitations que dans la parole conversationnelle [Vasilescu 2004] et une meilleure prédictibilité par l'évaluateur. Cependant, bien qu'il existe une pléthore de textes dans différentes langues, le constat a été fait dans le travail de thèse de Timothy Pommée tirée de [Pommée 2021a] qu'aucun d'entre eux ne semble vraiment répondre aux

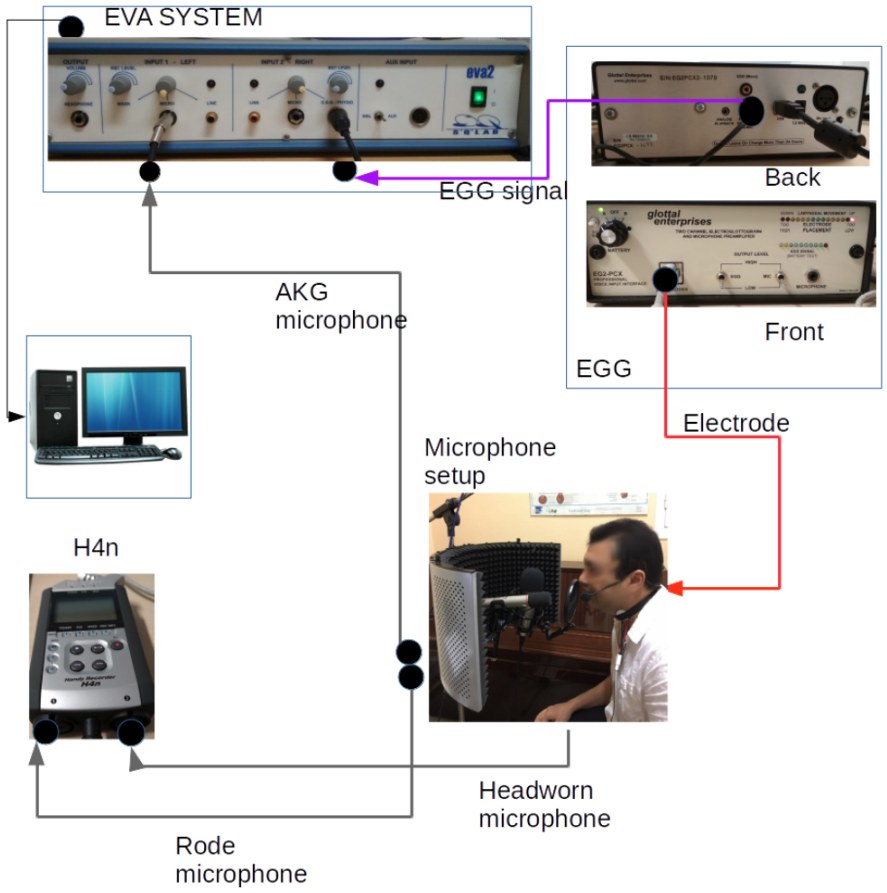

FIGURE 3.2 – Représentation du protocole d'enregistrement du corpus Voice4PD-MSA

attentes et aux besoins des cliniciens et des chercheurs pour l'évaluation courante de la parole et de la voix. Le texte a été construit sur la base d'un ensemble exhaustif de critères, prenant en compte les données de la littérature, les besoins spécifiques identifiés en recherche scientifique et en pratique clinique francophone, ainsi que les données d'une étude de consensus internationale. Ceci (voir figure 3.3) s'est fait en trois grandes étapes :

1. Recensement des besoins des cliniciens et chercheurs (mesures à appliquer sur le futur texte, population cible) et définition de l'objectif du texte ;
2. Recensement de critères de construction, sélection et hiérarchisation ;
3. Construction du nouveau texte.

L'article [Pommée 2024] donne plus d'informations sur la constitution de ce nouveau texte présenté dans l'encadré 3.2.3.

Le texte « Le voyage d'Alice » est destiné à fournir un support standardisé pour l'évaluation de l'articulation des sons de la parole (dysarthrie, apraxie), des variations prosodiques et du comportement phonatoire (dysphonie, harmonisation vocale), ainsi que de la fluence/des disfluences (bégaiement/bredouillement), chez les patients âgés d'au moins 12 ans. Il s'agit d'un outil adapté à la fois

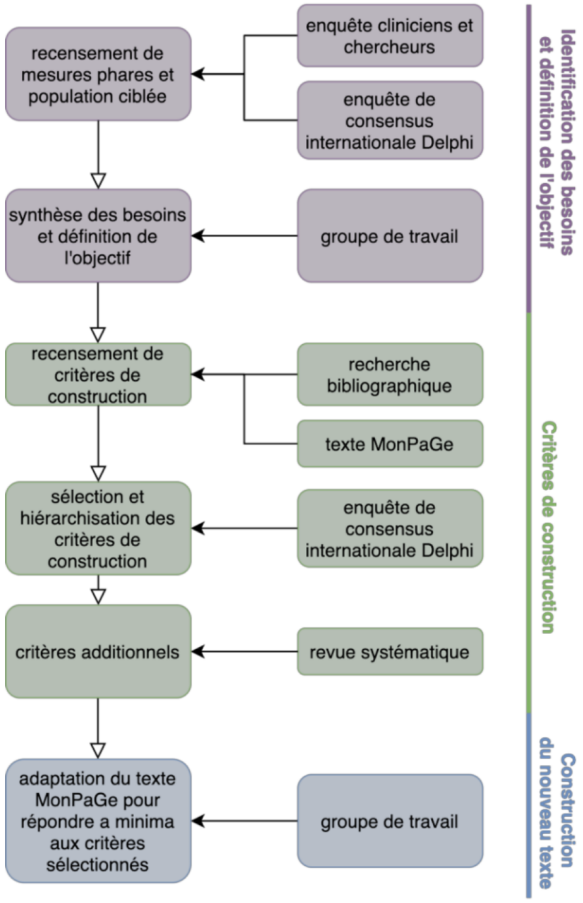

FIGURE 3.3 – Processus de création du nouveau texte pour l'évaluation de la parole et de la voix tiré de [Pommée 2021a]

pour la recherche scientifique, mais aussi potentiellement pour la pratique clinique quotidienne. Le protocole automatisé d'extraction de mesures acoustiques permet une analyse gratuite et rapide de données reproductibles, sans nécessiter une connaissance approfondie en informatique. Le tutoriel disponible en ligne devrait permettre une prise en main brève et simple de cet outil. Une fois l'installation initiale effectuée, le processus d'analyse qui se fait de manière asynchrone est entièrement automatisé et prend moins de 10 minutes.

Après la constitution de corpus, un travail d'annotation est à faire. Pour cela, plusieurs paradigmes peuvent être envisagés et plusieurs outils sont mis à disposition, notamment lorsqu'il s'agit au clinicien d'évaluer perceptivement la voix ou la parole et de donner une référence aux enregistrements composant le corpus. La partie suivante s'intéresse aux outils d'évaluation déjà disponibles en milieu clinique. Répondent-ils aux besoins des cliniciens ? Quelles sont leurs limites ?

Encadré 3.2.3 – Transcription orthographique du texte « Le voyage d'Alice » . La partie en gras correspond à une possibilité d'effectuer une passation rapide

Lundi matin, Alice et son Papa vont à Malibou.
Là-bas, ils rejoignent Papy après un voyage sans soucis.
Il fait chaud, mais la brise légère et l'air iodé de la mer les ravivent.
Vers midi, Alice s'exclame : « J'ai vraiment très très faim ! ».
Papy les guide alors vite vers un café luxueux au bord de l'eau : Le Bigorneau Salé.
Mardi, ils vont à la plage.
Il n'y a pas un nuage dans le ciel. Papa s'interroge : « Avons-nous pris la crème solaire ? »
« Bien sûr ! », répond Alice.
Mercredi, Papa et Papy se baladent en bavardant.
Pendant ce temps, Alice se détend en lisant un roman et mange un bonbon à l'ananas.
Jeudi, elle va faire un jogging.
Papa lui crie : « Nous partons faire quelques achats ! »
Au magasin, Papy achète des noix de macadamia.
Vendredi, ils visitent un musée d'art abstrait.
Papa s'extasie devant un splendide tableau et demande : « Qui a donc créé cette œuvre ? ».
Samedi matin, Alice s'entraîne pour la soirée karaoké en répétant rapidement : « pataka pataka pataka ».
Samedi soir, ils fêtent leur départ en dansant la java sous le lilas.
Comme à l'arrivée, il fait chaud, mais la brise légère et l'air iodé de la mer les ravivent.
Dimanche, Alice, Papa et Papy quittent Malibou.
Ils rentrent affamés.
À table, il y a de la pizza garnie et des lasagnes aux champignons.
Rassasiés, ils s'exclament : « Quel séjour extraordinaire ! »

3.3 État des lieux des outils d'évaluation clinique disponibles

Comme décrit dans la section 1.3.3, une vérité terrain est nécessaire pour étiqueter la parole des patients. Cette annotation peut se faire si un corpus de voix est disponible et si un jugement perceptif est réalisé sur ces enregistrements de voix. Pour ce faire, l'évaluateur peut s'appuyer sur les batteries de tests déjà présentes en consultations thérapeutiques, notamment lors du bilan de parole. Les méthodes d'évaluations disponibles en milieu clinique seront l'objet de la première partie de ce chapitre. Elles permettent d'analyser différents concepts globaux (p. ex. la sévérité du trouble, le caractère naturel, la clarté de la parole) ou plus spécifiques (p. ex. la prosodie, la qualité vocale, le débit de parole). Elles peuvent s'appliquer sur des unités de production de la parole variées, allant du phonème au texte ou à la parole semi-spontanée sur une échelle de granularité représentée figure 3.4. Chaque unité permet une évaluation répondant à un objectif spécifique, qui peut différer selon la nature de l'évaluation - perceptive ou instrumentale.

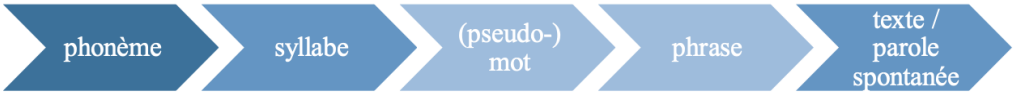

Figure 3.4 – Échelle de granularité des unités de production de la parole, tirée de [Pommée 2021a]

Enfin, des questionnaires d'auto-évaluation sont parfois utilisés pour évaluer la perception du trouble de la parole par le patient, ainsi que l'impact fonctionnel et la qualité de vie.

Comme indiqué dans le chapitre 3 de la thèse de Timothy Pommée [Pommée 2021a], au vu de la diversité des méthodes d'évaluation de la parole, mais aussi de la variabilité internationale de la formation des orthophonistes, les pratiques d'évaluation doivent être explorées dans les différents pays et communautés linguistiques. Alors que des enquêtes ont été menées dans différents pays et communautés linguistiques afin de développer des lignes directrices cliniques et mieux comprendre les attentes [Collis 2012, Conway 2015, Gurevich 2017, Miller 2017, Rumbach 2019], la littérature actuelle ne fournit pas de données sur les pratiques d'évaluation de la parole dans les pays francophones. Dans [Pommée 2022], la description d'une enquête destinée aux orthophonistes et phoniatres francophones est décrite, le premier objectif de cette étude étant de fournir un aperçu des pratiques cliniques actuelles concernant l'évaluation des troubles de la parole chez les patients adultes. Cette vue d'ensemble a ensuite permis d'aborder le second objectif de cette étude, à savoir l'identification des manques et des besoins rapportés par ces cliniciens concernant l'évaluation de la parole chez l'adulte. Cette enquête a d'abord consisté en le recrutement d'orthophonistes et phoniatres francophones via plusieurs moyens décrits dans [Pommée 2022]. Ces participants ont dû répondre à un questionnaire en ligne réalisé à l'aide de la plateforme LimeSurvey. Le questionnaire comprenait 49 questions : 18 questions ouvertes (entrées numériques, textes courts et textes longs) et 31 questions fermées (choix unique et multiple, réponses binaires, classements, échelles de Likert), regroupées en six catégories principales :

1. Informations sur le participant ;
2. Données sur le parcours académique et professionnel du participant et sur sa pratique actuelle, afin de pouvoir analyser les pratiques d'évaluation en fonction des différents profils de cliniciens ;
3. Informations concernant l'expérience professionnelle dans le domaine des troubles de la parole, afin d'estimer le niveau d'expertise du répondant ;
4. Des questions sur la patientèle, pour décrire le groupe cible des évaluations de la parole ;
5. Données sur l'équipement matériel/logiciel et l'évaluation de la parole effectuée pour les troubles de la parole, afin d'examiner les pratiques actuelles ;
6. Informations sur les éventuelles lacunes de l'évaluation de la parole.

Certaines questions étaient facultatives, afin d'éviter l'abandon du questionnaire en cas de difficultés de réponse. Une case « commentaire » était prévue pour permettre aux participants de compléter leurs réponses si nécessaire.

La figure 3.5 représente la distribution des niveaux de satisfaction des personnes interrogées concernant les outils d'évaluation de la parole. Dans l'ensemble, les outils disponibles ne sont jugés que moyennement satisfaisants.

Comme indiqué dans la thèse de T. Pommée, les commentaires des orthophonistes complètent ceci et soulignent le manque de fiabilité des outils d'évaluation actuellement disponibles, qui sont largement basés sur des évaluations subjectives. En voici quelques exemples :

— « J'ai l'impression de faire un bilan davantage qualitatif que quantitatif, malgré l'utilisation de la batterie ECD » ;
— « La cotation, même si elle se veut objective, demeure fonction de l'examinateur, de sa sensibilité et de ce qu'il considère comme étant pathologique (ortho expérimentée vs stagiaire ou ortho débutante par exemple) » ;
— « Fiabilité longitudinale si on garde les mêmes méthodes d'enregistrement et le même juge (or biais dans l'évaluation), fiabilité interjuges délicate. » ;

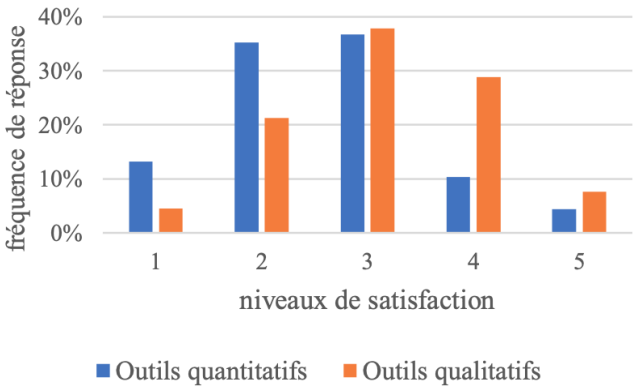

FIGURE 3.5 – Distribution des niveaux de satisfaction attribués aux outils d'évaluation de la parole, tirée de [Pommée 2021a]

— « Il y a selon moi une part assez importante à la subjectivité du praticien dans la notation des résultats du patient concernant la réussite de certains items parfois » ;
— « Bilans absolument pas reproductibles donc fiables ! On s'est basés sur la BECD pour une étude clinique, c'est une catastrophe en termes de reproductibilité et tous nos résultats sont biaisés ! » ;
— « Manque d'outils objectifs notamment sur la sévérité globale du trouble de la parole et sur la quantification des altérations segmentaires » .

Devant ce manque de fiabilité, la figure 3.6 tente de schématiser les manques rapportés et les solutions souhaitées par les cliniciens en trois grandes catégories :

— le manque de validité et de fiabilité des outils (en orange dans la Figure 3.6) : il s'agit principalement de la subjectivité de ces outils, entraînant un manque de fiabilité et de reproductibilité, ainsi qu'un manque de repères normatifs ;
— les difficultés liées aux aspects pratiques de l'utilisation des outils (en rouge dans la Figure 3.6) : de nombreux cliniciens se plaignent, entre autres, de la nature chronophage et onéreuse de ces outils et de la complexité de leur accès et de leur utilisation ;
— des lacunes dans l'applicabilité clinique (en jaune dans la Figure 3.6) : les outils actuellement disponibles manquent d'exhaustivité, sont insuffisants pour la pose de diagnostic et la planification de la prise en charge, et ne permettent pas l'évaluation de paramètres spécifiques tels que la prosodie, la nasalité et le débit de parole ; les outils d'évaluation quantitative fiables semblent rares à ce jour, et les thérapeutes y sont insuffisamment familiarisés.

Comme rapporté dans la thèse de Timothy, en ce qui concerne le manque de validité et de fiabilité, d'après les solutions souhaitées rapportées par les cliniciens, ces derniers profiteraient de nouvelles mesures reproductibles, ainsi que de critères de cotation clairement définis et de bases de données audio de référence pour améliorer les évaluations subjectives, et de données normatives pour comparer les performances des patients. Toujours selon les commentaires des répondants, les aspects pratiques de l'utilisation d'un tel outil d'évaluation seraient améliorés par le développement d'une solution abordable, facile d'accès comme d'utilisation, avec des conditions simplifiées d'enregistrement et de stockage des données. La portabilité serait également un grand avantage dans les différents contextes cliniques. Afin de pallier le manque d'exhaustivité des outils actuellement existants, nos résultats montrent qu'une attention particulière

devrait également être accordée à certains paramètres tels que la prosodie, la nasalité et le débit de parole. De plus, des tâches spécifiques, telles que l'utilisation de pseudo-mots pour cibler les paramètres de bas niveau de la parole, devraient être combinées avec des tâches d'évaluation plus écologiques. En outre, des questionnaires adressés à l'entourage du patient pourraient permettre d'obtenir une vision plus complète de la communication du patient. Enfin, les cliniciens bénéficieraient également d'une meilleure information sur les outils d'évaluation de la parole existants et des formations continues, notamment en ce qui concerne l'évaluation acoustique des troubles de la parole.

Une fois fixée sur l'outil d'évaluation à adopter, les termes définis, le pool d'évaluateurs et le corpus à utiliser, la partie caractérisation automatique des enregistrements audio peut démarrer.

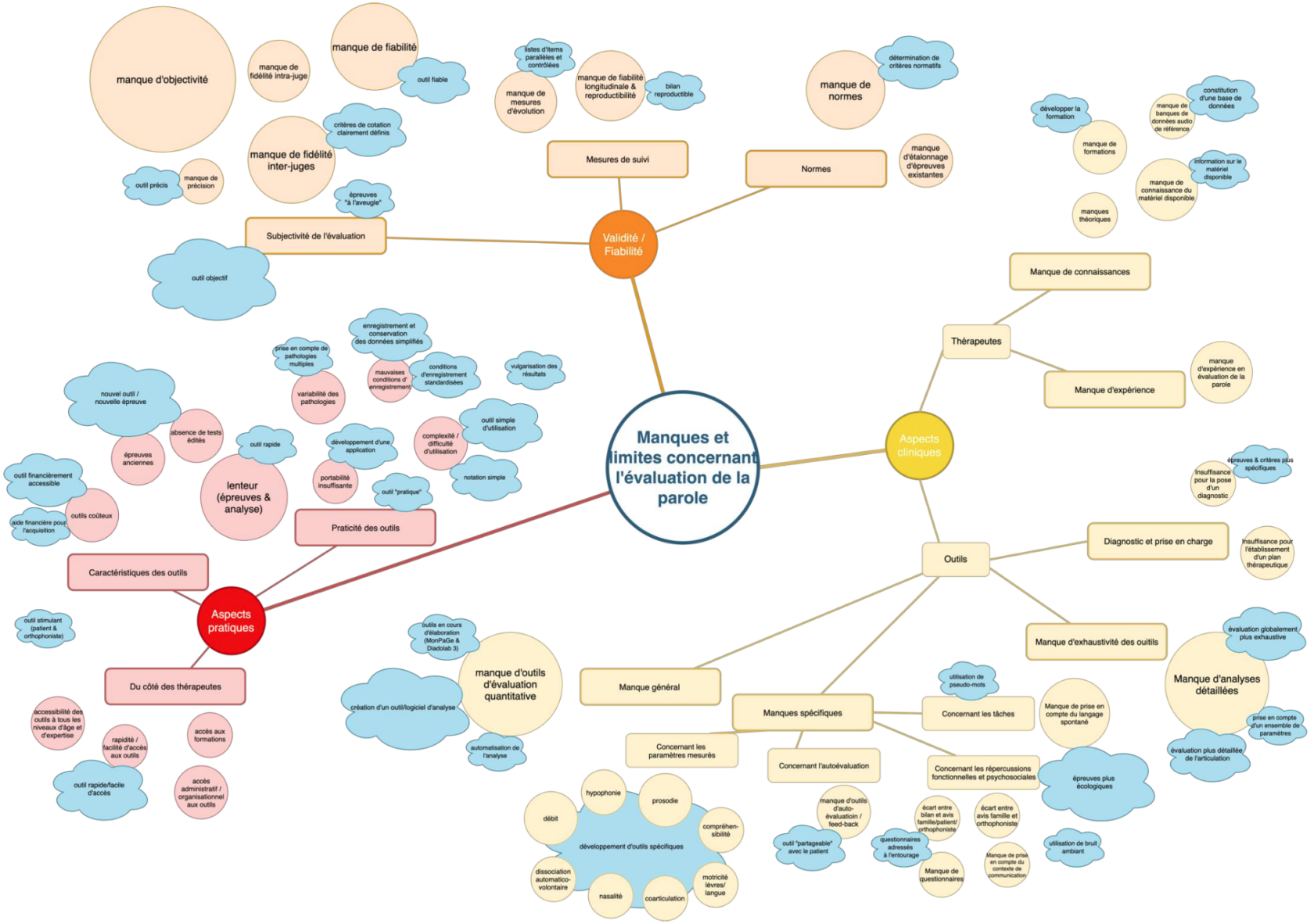

FIGURE 3.6 – Manques rapportés et solutions souhaitées par les cliniciens concernant l'évaluation des troubles de la parole, tirée de [Pommée 2021a]

3.4 Caractérisation de la parole pathologique

3.4.1 Les pauses silencieuses et les pauses remplies dans les troubles aphasiques

Les perturbations de la fluence, telles que les pauses, les faux départs, les auto-corrections, l'utilisation abusive de pauses remplies, et les répétitions, sont fréquentes dans la production orale des individus atteints d'aphasie non fluente. Caractériser ces perturbations peut permettre de mieux comprendre les différentes stratégies mises en place par les locuteurs atteints de cette maladie.

Dans les deux études décrites par la suite, la méthodologie adoptée a été élaborée durant deux stages de M1 que j'ai co-encadrés avec Halima Sahraoui, ceux de Antonin Klopp-Tosser et Victor David. Le premier stagiaire a observé les pauses remplies et silencieuses grâce à un traitement automatique fourni par le logiciel Easyalign [Goldman 2011]. Le deuxième a amélioré la détection des pauses remplies et des allongements vocaliques en combinant l'algorithme de divergence Forward-Backward [André-Obrecht 1988] pour détecter les zones quasi stationnaires du signal audio, combinée avec la méthode de segmentation en pseudo-syllabe [Farinas 2002] permettant de regrouper les voyelles « allongées » qui auraient été séparées en plusieurs segments.

Dans l'étude [Sahraoui 2015], nous avions ainsi formulé l'hypothèse selon laquelle les variations de fluence verbale en production de discours continu sont inversement corrélées aux variations de précision/complexité morpho-syntaxique, et ce en raison de stratégies attentionnelles de contrôle de la production (monitoring) plus caractéristiques dans certaines tâches langagières où le locuteur recherche plus de précision grammaticale. Les tâches étudiées ici sont le discours autobiographique du patient et la description d'images. Ainsi, afin d'investiguer cette question de recherche, il fut nécessaire de mieux décrire les variations de fluence verbale dans un tel trouble aphasique, qualifié classiquement de « non fluent » , et donc de réaliser des analyses fines de corpus (pathologiques et non pathologiques à titre de comparaison avec la performance normale). L'idée était de se donner les moyens pour combiner des analyses structurales syntaxiques, lexicales, morphologiques (élaboration du discours) avec des analyses des indices prosodiques, en particulier les durées et places des pauses, le nombre de reformulations et d'hésitations, l'intonation. Nous avons pu mieux comprendre l'organisation des pauses silencieuses et des pauses remplies en discours, leur rôle, et le fait que les pauses silencieuses, en particulier, reflèteraient des processus de contrôle ou de planification de la production orale en discours.

Les stratégies correctives ne sont possibles que pour les patients ayant une capacité préservée à détecter les erreurs et à surveiller leur production de la parole [Postma 2000, Oomen 2001]. À la suite de cette première étude, nous avons formuler l'hypothèse dans [Sahraoui 2022] que les individus atteints d'aphasie non fluente, ayant généralement des déficits de compréhension mineurs, peuvent surutiliser les compétences de surveillance dans la production du langage à un stade pré- ou post-articulatoire et que la surveillance peut varier en fonction du type de tâche [Sahraoui 2015].

3.4.2 Étude du comportement du Goodness of Pronunciation sur la parole pathologique

Dans ce paragraphe nous rapportons une étude que nous avons effectuée visant à détecter automatiquement les erreurs de prononciation au niveau des phonèmes chez 32 locuteurs français souffrant de paralysie faciale unilatérale (voir paragraphe 3.2.1), classés en quatre grades de gravité clinique diffé-

rents [Pellegrini 2014]. Les locuteurs de grade I de l'échelle de House et Brackmann sont dans le groupe contrôle puis les locuteurs des grade V et VI sont regroupés. Nous avons cherché à déterminer si l'algorithme Goodness of Pronunciation (GOP), décrit dans le paragraphe 2.5 pouvait également détecter les déviations segmentales dans la parole pathologique. À cette fin, la parole lue par les 32 locuteurs a été alignée et des scores GOP ont été calculés pour chaque réalisation de phonème. Les scores les plus élevés, indiquant de grandes dissimilarités avec les réalisations standard des phonèmes, ont été obtenus pour les locuteurs les plus gravement atteints. Le sous-ensemble de parole correspondant a été transcrit manuellement au niveau des phonèmes. 8,3% des phonèmes différaient des prononciations standard extraites de notre lexique. La technique GOP a permis de détecter 70,2% des erreurs de prononciation avec un taux d'égale erreur d'environ 30% de faux rejets et de fausses acceptations. Les substitutions de phonèmes détectées par l'algorithme ont confirmé que certains locuteurs ont des difficultés à produire des occlusives bilabiales et ont montré que d'autres sons, tels que les sifflantes, sont sujets à des erreurs de prononciation. Par exemple, la figure 3.7 montre les substitutions croissantes des [s] des patients.

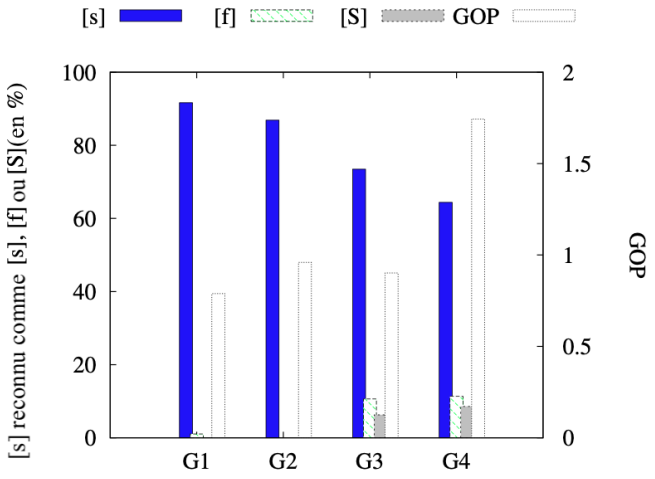

FIGURE 3.7 – Moyenne des scores GOP et substitutions les plus fréquentes pour le [s], d'après la reconnaissance automatique tirée de [Pellegrini 2014]

Comme on peut le voir, les confusions avec [f] et [S] augmentent avec le grade de l'atteinte, tout comme le GOP moyen pour ce phonème, à l'exception du grade G3. Néanmoins, l'évolution du score GOP n'était pas toujours aussi claire pour d'autres phonèmes, et les locuteurs diagnostiqués avec un même grade de pathologie ne partagent pas nécessairement les mêmes problèmes de prononciation. De plus, les scores GOP moyens (voir tableau 3.2) ne corrèlent pas fortement avec les grades de gravité clinique, mesurés par l'échelle de House-Brackmann. Selon nous, ce résultat pourrait refléter que non seulement l'échelle H&B a été conçue pour évaluer la mobilité faciale globale - c'est-à-dire pas uniquement celle des articulateurs de la parole et que certains patients peuvent employer des stratégies compensatoires efficaces afin de rester intelligibles malgré les déficiences motrices dont ils souffrent. Dans cette perspective, un outil automatique tel que le GOP constituerait un moyen intéressant pour collecter facilement des données pertinentes concernant la capacité de communication des patients.

TABLEAU 3.2 – Moyenne et écart type des scores GOP pour les différents groupes de patients

Groupes de gravité	Moyenne du score GOP (écart type)
G1	1.68 (2.98)
G2	1.94 (3.12)
G3	1.72 (2.86)
G4	2.25 (3.50)

3.4.3 Analyse de la parole pour le diagnostic différentiel entre le Maladie de Parkinson et l'Atrophie Multi-Systématisée

Dans ce chapitre est exposée l'étude et la recherche de marqueurs issus d'un traitement automatique de la parole, afin d'aider un diagnostic différentiel entre deux maladies, la Maladie de Parkinson (MP) et l'Atrophie Multi-Systématisée (AMS) [Laaridh 2020].

Cette étude a été faite sur le corpus Voice4PD décrit au paragraphe 3.2.3 et plus spécialement sur la tâche de lecture du texte « La chèvre de monsieur Seguin » d'environ 70 mots. Afin de pouvoir identifier le potentiel et la pertinence d'indices acoustiques de manière la plus objective possible, nous avons privilégié leur extraction suivant la méthodologie du GOP, cette fois-ci en utilisant la boîte à outils KALDI [Povey 2011] :

— Le premier est un alignement forcé du signal de parole sur la suite phonétique contrainte par le texte prononcé.
— Le deuxième est une tâche de reconnaissance automatique du signal en une suite de phonèmes, sans connaissance a priori du texte prononcé et sans lexique.

Les indices acoustiques étudiés sont extraits des sorties de chacun de ces traitements ; ils peuvent ensuite être analysés indépendamment du traitement ou comparés.

Dans cette publication [Laaridh 2020], Imed Laaridh et moi proposons trois indicateurs issus du prétraitement basé sur la simple reconnaissance automatique de phonèmes, à savoir la qualité de la reconnaissance phonétique, la durée des voyelles reconnues et le débit de parole. Des différences significatives dans le calcul de ces paramètres sont en adéquation avec les observations effectuées par le corps médical sur les problèmes de prononciation dus à l'AMS, à savoir une qualité acoustique moindre, des voyelles plus longues et un débit de parole plus lent.

Qualité acoustique au niveau phonème

L'alignement forcé sur le texte lu tente de localiser tous les phones au risque de pénaliser par endroit la qualité acoustique. À l'inverse, le système de reconnaissance phonétique est basé entièrement sur la reconnaissance d'une information acoustique indépendante du locuteur. Une approche pour mesurer la qualité de la réalisation acoustique de chaque enregistrement consiste à aligner temporellement les sorties de l'alignement contraint par le texte et de la reconnaissance automatique phonétique, et à comparer les phones ainsi associés trame à trame (toutes les 10 ms). Le pseudo taux de reconnaissance phonémique, à savoir le ratio de trames bien reconnues par la reconnaissance automatique, s'apparente à un indicateur de qualité TR :

$$TR = 100 * \frac{\#\ trames\ bien\ reconnues\ par\ la\ reconnaissance\ automatique}{\#\ de\ trames\ de\ référence} \tag{3.1}$$

Les mesures de l'indicateur *TR* pour chaque individu des trois populations du corpus d'étude sont rapportées sur la figure 3.8 : un point représente un individu. Les trois populations sont isolées les unes des autres afin de pouvoir visualiser la moyenne et la variance pour chacune d'elles. Plus ce taux est important, mieux la parole a été reconnue par rapport à une réalisation acoustique standard.

Pour la population « témoin » , le taux de reconnaissance phonémique est de 62%, un taux normal pour un système indépendant du locuteur, ce qui valide la qualité des modèles acoustiques appris et justifie l'examen des résultats sur les deux autres populations.

Pour les patients atteints d'AMS, le système a beaucoup plus de difficulté à reconnaître les phonèmes prononcés : seulement 38% de trames sont bien reconnues sur l'ensemble des enregistrements contrairement aux patients atteints de la MP pour qui le *TR* moyen est de 57%, proche du groupe « témoin » . Une différence significative entre ces deux pathologies est confirmée statistiquement ($p<0{,}001$) (Anova à un facteur). Ce comportement reflète une information importante que nous retrouvons dans la littérature caractérisant ces deux pathologies : la dysarthrie liée à l'AMS est souvent plus sévère que celle associée à la MP. Cette sévérité est reflétée dans la qualité de la reconnaissance automatique des phonèmes et l'indicateur *TR* présente un potentiel pour son utilisation pour la différenciation entre les deux pathologies.

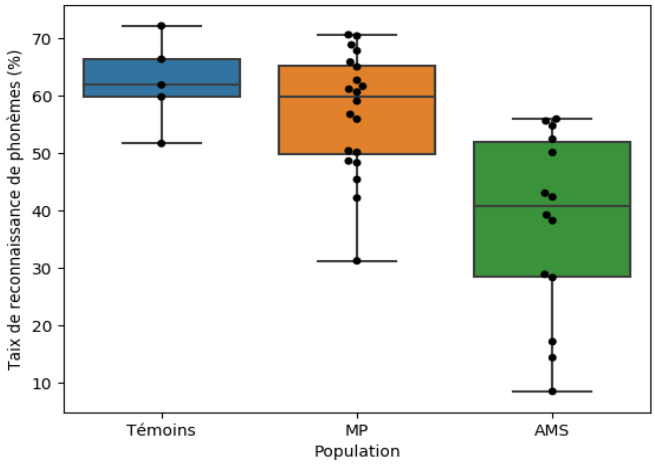

FIGURE 3.8 – Valeur de l'indicateur *TR* (%) par individu et par population

Durée moyenne des voyelles

L'allongement des phonèmes, plus particulièrement celui des voyelles, est souvent associé à la dysarthrie ataxique [Darley 1975]. Vu que la dysarthrie mixte associée à l'AMS comporte une composante ataxique, absente de la dysarthrie associée à la MP, l'étude de la durée moyenne des voyelles chez les patients des deux populations est pertinente. Ainsi, l'indicateur Dur_{voy} est obtenu en moyennant sur chaque enregistrement, les longueurs des segments reconnus comme une voyelle :

$$Dur_{voy} = \frac{Durée\ totale\ des\ voyelles}{\#\ de\ voyelles} \tag{3.2}$$

Cet indicateur Dur_{voy} est calculé sur chaque enregistrement à l'issue de la reconnaissance automatique de phonèmes, des travaux antérieurs de segmentation automatique du signal de parole ayant montré que les frontières détectées par les systèmes de reconnaissance phonétique sont fiables, même si l'identification phonémique en elle-même peut être erronée.

Les résultats sont rassemblés sur la figure 3.9. Les patients atteints d'AMS présentent un comportement distinctif, avec des prononciations de voyelles significativement plus longues que la normale ($p<0{,}001$). Ce résultat confirme la pertinence de la mesure Dur_{voy} qui reflète l'allongement des voyelles associé à la dysarthrie ataxique, comportement décrit dans la littérature et basé sur l'analyse perceptive de la parole.

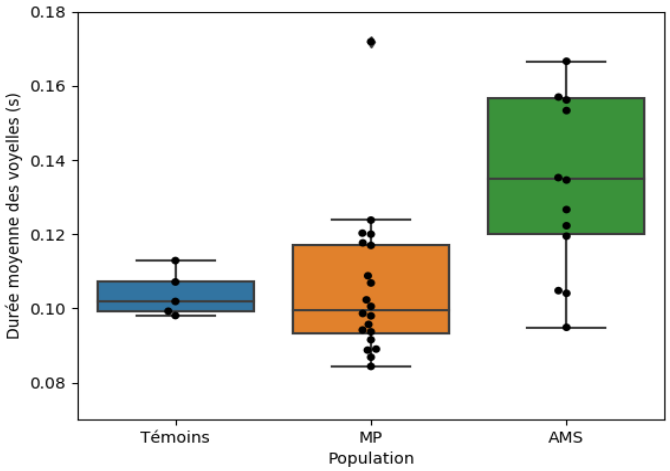

FIGURE 3.9 – Dur_{voy} par population à partir de la reconnaissance automatique de phonèmes

Débit de parole

L'allongement de phonème observé chez les patients atteints de l'AMS agit a fortiori sur leur débit de parole. La dysarthrie ataxique dont souffrent ces patients est caractérisée par un ralentissement du débit de la parole, contrairement à la dysarthrie hypokinétique associée à la MP caractérisée par des irrégularités (voire des accélérations) du débit.
Afin d'estimer automatiquement le débit de parole, nous proposons de calculer le nombre de voyelles prononcées par seconde, noté *Débit* ; c'est un estimateur proche du nombre de syllabes par seconde et de la vitesse d'élocution, par ailleurs couramment utilisé [Rouas 2004].

$$Débit = \frac{\#\ de\ voyelles}{Durée\ totale\ de\ lecture - (Durée\ pause\ + Durée\ respiration)} \tag{3.3}$$

Le paramètre *Débit* est extrait à partir des sorties de la reconnaissance automatique de phonèmes. Pour éviter tout biais, les pauses et respirations ne sont pas prises en compte pour le calcul.

Les résultats par type de population sont rassemblés sur la figure 3.10. Comme pour l'étude de la durée des voyelles, des résultats prometteurs s'observent : les patients atteints d'AMS présentent un débit de parole significativement plus lent que les locuteurs témoins ou atteints de la MP ($p < 0.01$).

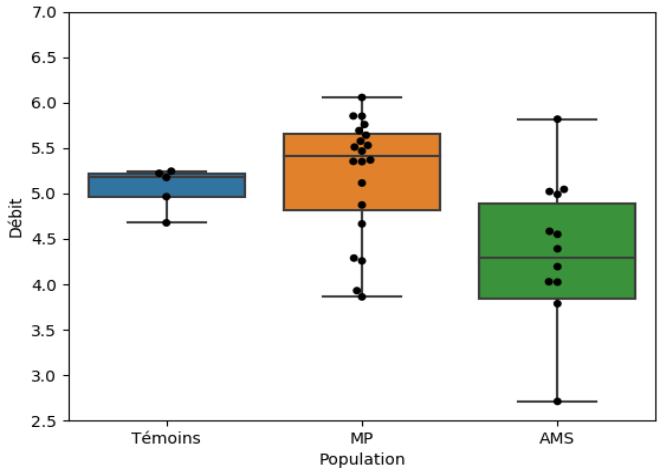

FIGURE 3.10 – *Débit* par population à partir de la reconnaissance automatique de phonèmes

La variance élevée de ce paramètre sur chaque population peut s'interpréter de deux manières différentes : soit les individus au sein d'une même population présentent des comportements très différents l'un de l'autre (plus que probable), soit chaque individu a un comportement très irrégulier au cours du temps et le paramètre, *Débit*, moyenne sur l'ensemble du texte lu, n'est pas significatif, compte tenu de l'instabilité éventuelle d'un débit « local » .

Des différences significatives avec MP dans le calcul de ces paramètres sont en adéquation avec les observations effectuées par le corps médical sur les problèmes de prononciation dus à l'AMS, à savoir une qualité acoustique moindre, des voyelles plus longues et un débit de parole plus lent. Ceci laisse espérer des résultats encourageants sur une aide au diagnostic précoce entre MP et AMS. À noter également que le calcul de ces paramètres ne demande aucune connaissance a priori qu'une transcription phonétique du texte produit par les patients ; il ne nécessite pas de travail supplémentaire d'annotation ou de transcription sur le signal.

Après cette partie sur la caractérisation acoustique utilisée sur plusieurs corpus de parole pathologique à des fins d'aide au diagnostic, la prochaine partie traite de modélisation de parole pathologique dans le cadre de l'intelligibilité de la parole de patients ayant des troubles de la communication. La question de recherche sous-jacente ayant été l'intitulé d'une des bourses du projet TAPAS est de savoir si une modélisation par réseaux de neurones profonds est à même de fournir un score d'intelligibilité de la parole d'un patient.

3.5 Modélisation de la parole pathologique

Le dernier paragraphe de ce chapitre est consacré aux travaux de thèse de Sebastião Quintas [Quintas 2022a] que j'ai co-encadré autour de la modélisation par réseaux de neurones profonds. L'application visée est l'intelligibilité de la parole avec notamment la question de recherche centrale : *l'apprentissage profond peut-il être utilisé de manière fiable pour prédire l'intelligibilité de la parole ?*

Malgré les avancées récentes en apprentissage profond, le sujet de l'interprétabilité est souvent négligé,

principalement parce que certains domaines d'application ne nécessitent pas une justification approfondie des scores obtenus. Bien que les résultats dans certains de ces domaines puissent parler d'eux-mêmes et ne nécessitent pas une réflexion approfondie, cela ne s'applique pas nécessairement dans le cas de la parole pathologique. Un certain degré d'interprétabilité est toujours nécessaire notamment pour être plus accepté du corps médical et des patients. Les termes : interprétabilité/explicabilité seront définis par la suite comme le degré auquel un humain peut comprendre la cause d'une décision [Miller 2019]. Étant donné la nature complexe des systèmes d'apprentissage profond, l'utilisation de scores de prédiction intermédiaires ou de caractéristiques significatives et cliniquement valides peut fournir cette couche supplémentaire d'explicabilité. Ces approches peuvent être considérées comme plus interprétables par rapport, par exemple, aux systèmes entièrement de bout en bout.

Dans un contexte automatique, il devient pertinent de ne pas se limiter à une seule approche ou mesure, surtout lorsqu'il existe plusieurs méthodes cliniques pour évaluer perceptuellement l'intelligibilité de la parole. Le processus de décodage acoustico-phonétique, décodage des unités de parole suivi de la conversion en graphèmes, est l'une des principales approches utilisées pour évaluer l'intelligibilité et peut être considéré comme plus objectif que l'évaluation subjective sur une échelle. Les unités de parole de ce processus peuvent être de différentes natures selon le test, allant des phonèmes et des syllabes aux mots, voire aux phrases complètes. Étant donné que chacun de ces niveaux différents est pertinent à sa manière (par exemple, l'analyse phonémique peut cibler des erreurs clés de prononciation tandis que l'analyse de phrases peut évaluer la communication quotidienne), une **approche granulaire** ciblant ces niveaux distincts devient très intéressante dans le contexte d'une analyse automatique. C'est cette partie du travail de Sebastião que je vais reprendre ici, en explorant la **prédiction automatique de l'intelligibilité de la parole utilisant les niveaux de : phrase, mot et phonème**.

3.5.1 Modélisation au niveau de la phrase

Cette partie du travail de Sebastião a été publié dans [Quintas 2020]. La méthodologie proposée repose sur deux étapes (voir figure 3.11). La première correspond à l'extraction des *x-vecteurs* des locuteurs [Snyder 2018a], afin d'obtenir une représentation de longueur fixe pour chaque énoncé de locuteur. Des travaux récents suggèrent en effet que les *x-vecteurs* peuvent être appliqués avec succès à des tâches paralinguistiques telles que la reconnaissance des émotions [Pappagari 2020], ainsi qu'à la détection de maladies comme la maladie d'Alzheimer [Zargarbashi 2019]. Dans cette étude, nous avons utilisé les enregistrements segmentés d'une tâche de lecture de texte (LEC) qui est décrite dans 3.2.2. La deuxième étape repose sur une tâche de régression pour prédire un score d'intelligibilité basé sur les représentations extraites précédemment à l'aide d'un réseau de neurones peu profond.

Méthode de prédiction de l'intelligibilité

Comme mentionné précédemment, les *x-vecteurs* sont des représentations de caractéristiques de locuteurs DNN qui sont utilisées dans les tâches de reconnaissance de locuteurs et de paralinguistique [Pappagari 2020]. Alors que les *i-vecteurs* représentent le sous-espace de variabilité totale d'un canal ou d'un locuteur, les *x-vecteurs* visent à représenter des caractéristiques discriminantes entre les locuteurs. La comparaison des deux méthodes suggère que les *x-vecteurs* nécessitent des segments temporels plus courts pour obtenir de bons résultats, et se sont avérés plus robustes face à la variabilité des données et aux décalages de domaine [Snyder 2018b].

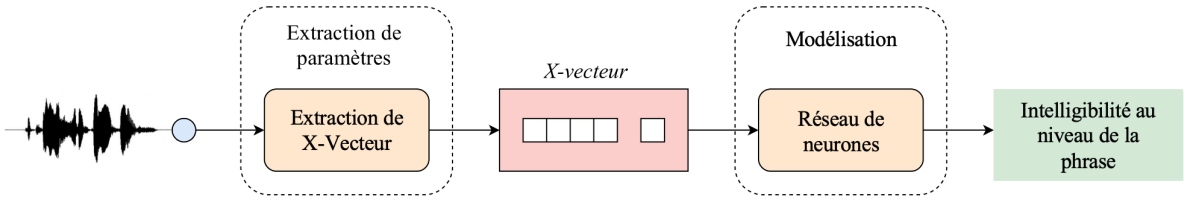

FIGURE 3.11 – Le système global : les x-vecteurs sont extraits de chacun des segments de phrases de la tâche de lecture (LEC), puis donnés en entrée d'un réseau de neurones qui prédit un score d'intelligibilité

Pour extraire les *x-vecteurs*, nous avons utilisé l'implémentation open source présente dans le toolkit Kaldi [1].

Tous les longs silences et les bruits ont été supprimés des fichiers audio d'entrée.

Ensuite, pour prédire un score d'intelligibilité basé sur les représentations x-vecteurs, un réseau neuronal peu profond a été modélisé. Seules des couches entièrement connectées ont été utilisées dans notre cas. Pour chaque locuteur, l'intelligibilité moyenne a été calculée sur la base de l'évaluation perceptuelle indépendante de six professionnels de santé différents. Chaque locuteur a reçu un score entre 0 et 10, plus la valeur est petite, moins la parole est intelligible.

Les enregistrements ont ensuite été segmentés en huit segments de longueur similaire, qui sont indiqués dans le texte. Un total de 105 locuteurs, dont 84 patients et 21 contrôles, ont été utilisés dans cette étude.

La tâche de lecture (LEC), a été découpée suivant les segments : (S_1) *Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres.* (S_2) *Il les perdait toutes de la même façon.* (S_3) *Un beau matin, elles cassaient leur corde,* (S_4) *s'en allaient dans la montagne, et là-haut le loup les mangeait.* (S_5) *Ni les caresses de leur maître* (S_6) *ni la peur du loup rien ne les retenait.* (S_7) *C'était paraît-il des chèvres indépendantes* (S_8) *voulant à tout prix le grand air et la liberté.*

Un schéma de validation croisée en 5 blocs a été mis en place pour entraîner le réseau neuronal peu profond. À chaque bloc, 84 locuteurs (patients et contrôles) ont été utilisés pour l'entraînement et les 21 locuteurs restants, non vus, ont été utilisés pour les tests. Pour chaque passe, le réseau neuronal peu profond a été entraîné pendant un total de 15 phases en utilisant une décroissance exponentielle du taux d'apprentissage. La normalisation par lot et un taux de dropout de 25% ont été appliqués à chaque couche.

Scores d'évaluation

Afin de juger les prédictions obtenues, nous avons évalué notre système selon deux métriques : la corrélation de Spearman (p), car les valeurs cibles d'intelligibilité étaient loin de suivre une distribution normale, et l'erreur quadratique moyenne (RMSE). Les scores ont été calculés en utilisant les valeurs perceptuelles mentionnées dans la section 3.2.2 comme référence. L'évaluation perceptuelle de l'intelligibilité est entièrement décrite dans [Astésano 2018].

1. https://github.com/kaldi-asr/kaldi

La première expérience de prédiction de l'intelligibilité que nous avons réalisée a utilisé les huit segments de chaque locuteur. Dans ce cas, les *x-vecteurs* ont été extraits, alimentés dans le réseau neuronal peu profond, puis une valeur d'intelligibilité est prédite. Le score final pour chaque locuteur a été calculé comme la moyenne des scores des huit segments de chaque locuteur. La figure 3.15 montre les valeurs d'intelligibilité prédites comparées à l'évaluation perceptuelle des professionnels. Les valeurs de corrélation obtenues sont cohérentes avec celles trouvées dans des études précédentes telles que [Laaridh 2018], qui ont obtenu des valeurs de corrélation entre 0,75 et 0,84. Cependant, il est important de noter que les mesures d'intelligibilité perceptuelle, dans notre cas, sont évaluées par des professionnels de la santé plutôt que par des auditeurs naïfs.

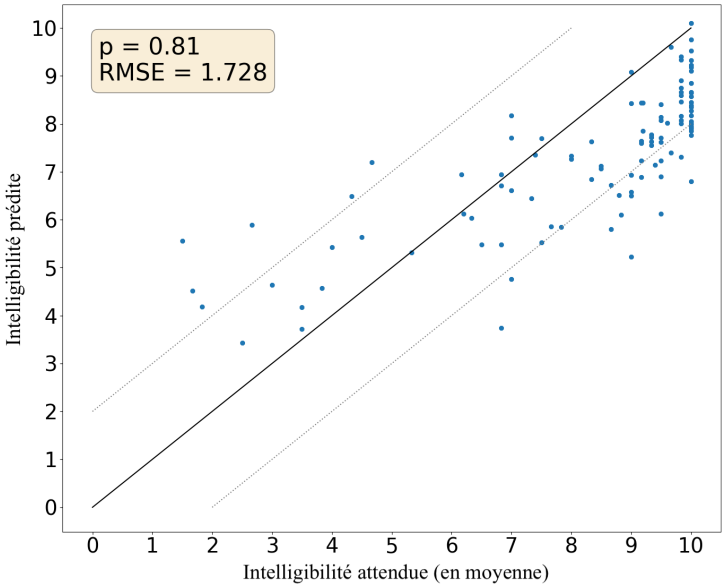

FIGURE 3.12 – Graphique de prédiction de l'intelligibilité, en utilisant la moyenne des scores des segments de chaque locuteur, tiré de [Quintas 2020]

À partir des résultats obtenus en utilisant le score moyen des 8 segments, nous avons remarqué que, dans la majorité des cas, il y avait une grande variance au sein des scores individuels de chaque locuteur. En analysant plus finement les résultats, on peut constater que certaines phrases permettent d'obtenir une estimation de l'intelligibilité beaucoup plus précise. Nous avons approfondi cet aspect en choisissant manuellement, pour chaque locuteur, le segment dont la valeur prédite était la plus proche de la valeur cible. Les valeurs de RMSE et de p ont été calculées. Les résultats peuvent être trouvés dans le tableau 3.3, accompagnés des résultats du choix du pire segment également. Les valeurs obtenues suggèrent que, pour chaque locuteur, certains segments sont capables de fournir une mesure d'intelligibilité très précise, affichant généralement une valeur de corrélation très élevée et une faible erreur quadratique moyenne.

Cela indique une analyse plus approfondie de ces segments, montrant que nous sommes capables d'obtenir des valeurs de corrélation très élevées en identifiant manuellement les énoncés qui conviennent le mieux à chaque locuteur. Les meilleurs segments résultants ont été évalués plus en détail. Bien qu'aucune préférence claire n'ait été trouvée pour une phrase spécifique, parmi la sous-liste des meilleurs segments, les segments numéro 2 et numéro 6 sont les plus représentatifs, représentant respectivement 23% et 15% des cas (voir figure 3.13).

TABLEAU 3.3 – Résultats obtenus en choisissant manuellement, pour chaque locuteur, le meilleur et le pire score de la phrase. La phrase moyenne présente les résultats obtenus en faisant la moyenne des huit phrases de chaque locuteur. Les meilleures et la pire phrase illustrent les résultats obtenus en choisissant manuellement, pour chaque locuteur, la phrase la plus proche et la plus éloignée de la cible.

	p	RMSE
Moyennes des phrases	0,81	1,728
Meilleure phrase	0,95	0,900
Pire phrase	0,53	2,224

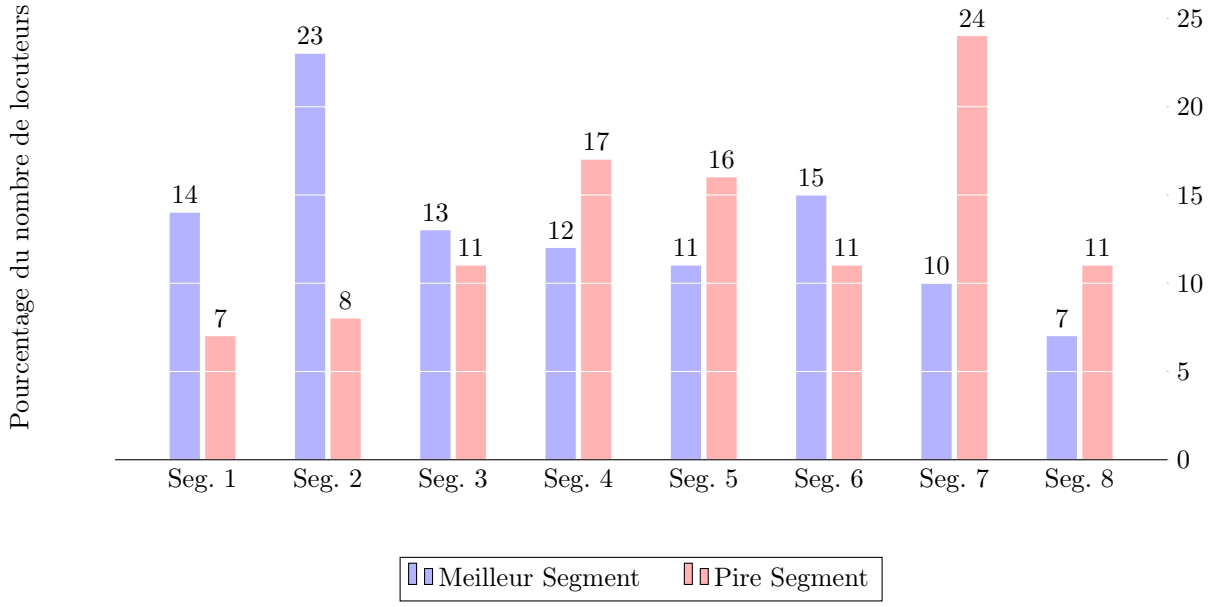

FIGURE 3.13 – Distribution de meilleures et pires phrases en pourcentage

Choisir la meilleure phrase

À partir du sous-ensemble de locuteurs pour lesquels les segments 2 et 6 prédisaient un score proche du score d'intelligibilité réel, nous avons analysé le nombre de phonèmes reconnus pour chaque locuteur de chaque sous-ensemble. En effet, les taux d'erreur de mots et de phonèmes sont depuis longtemps utilisés comme moyen automatique d'évaluer l'intelligibilité de la parole [Christensen 2012]. De cette analyse, nous avons trouvé que le segment 2 était associé à un taux de reconnaissance clairement supérieur à la moyenne du taux de reconnaissance, tandis que le segment 6 était nettement inférieur à la moyenne. Nous pensons que cela peut être dû à la présence plus importante de voyelles nasalisées dans le segment 6, qui présentent généralement des problèmes d'articulation pour les patients atteints de cancer de l'oropharynx [Jacobi 2013].

Plusieurs résultats intéressants sont observables dans le tableau 3.4. Tout d'abord, en utilisant simplement le segment 2, le couple corrélation/RMSE obtenu (p : 0,82, RMSE : 1,434) était légèrement meilleur que le score moyen affiché dans la tableau [**?**]. De plus, afin de choisir le meilleur segment automatiquement et non manuellement, nous avons conçu un simple arbre de décision qui choisit le meilleur segment pour chaque locuteur en fonction du nombre de phonèmes reconnus par un système de reconnaissance de

phonèmes basé sur Kaldi. Pour les locuteurs ayant au moins 167 phonèmes reconnus, le segment 2 a été assigné, tandis que le segment 6 a été choisi pour les autres. La valeur de 167 phonèmes reconnus mentionnée correspond à la valeur moyenne des phonèmes reconnus pour les patients de C2SI. Les résultats en utilisant ce critère, présentés dans le tableau 3.4, suggèrent une amélioration de la corrélation et une diminution totale de 0,34 de l'erreur quadratique moyenne.

TABLEAU 3.4 – Comparaison entre les scores obtenus précédemment, une baseline *i-vecteurs* obtenue grâce au modèle pré-entraîné [Povey 2011] et l'arbre de décision implémenté

		p	RMSE
i-vecteurs	Scores Moyens	0,72	2,121
x-vecteurs	Scores Moyens	0,81	1,728
	Segment 2 seul	0,82	1,434
	Arbre de Décision	0,85	1,389

Les résultats présentés suggèrent qu'en utilisant le paradigme des *x-vecteurs*, nous sommes capables d'obtenir des prédictions d'intelligibilité fiables au niveau de la phrase. De plus, en identifiant le meilleur segment / la meilleure phrase pour chaque locuteur, une valeur de corrélation très élevée peut être atteinte, et le RMSE diminue de presque moitié par rapport à la valeur obtenue avec l'approche moyennée.

3.5.2 Modélisation au niveau du mot

Il y a de nombreuses raisons de procéder à une évaluation de l'intelligibilité à ce niveau du mot. Au niveau de la phrase, les évaluations perceptives et automatiques tendent à adopter une approche plus globale de l'intelligibilité de la parole, en prenant en compte des caractéristiques suprasegmentales telles que la prosodie, la qualité de la voix, et la résonance. Cependant, lorsque l'on observe des unités plus petites, comme des mots, l'évaluation de ces paramètres devient difficile, en raison du manque de contexte temporel notamment. Par conséquent, une évaluation de l'intelligibilité au niveau des mots est souvent plus appropriée pour juger de l'articulation correcte ou incorrecte du patient.

Au niveau perceptuel, beaucoup d'approches sont utilisées pour évaluer l'intelligibilité de la parole. L'évaluation dans un contexte clinique a souvent une variabilité élevée, qui peut se traduire par des résultats différents donnés par le même praticien à travers différentes tâches. Ainsi, des alternatives comme le décodage acoustico-phonétique (noté DAP par la suite) peuvent être plus objectives et pertinentes non seulement pour le contexte clinique, mais aussi pour entraîner des systèmes automatiques [Ghio 2018]. Notre approche utilise des transcriptions faites par des auditeurs naïfs. L'objectif est de prédire automatiquement le score DAP, dans le contexte de patients atteints de cancers des voies aérodigestives supérieures. D'une part, nous souhaitons obtenir un système fiable. D'autre part, nous voulons évaluer l'impact de la quantité de données (nombre de pseudo-mots utilisés) sur la qualité des prédictions. Le score DAP, même s'il est obtenu par la transcription de pseudo-mots au lieu de l'évaluation perceptive classique, est considéré comme une mesure d'intelligibilité précieuse, avec des applications pratiques très concrètes et objectives dans un contexte clinique.

Dans le cadre des prédictions automatiques au niveau du mot de l'intelligibilité de la parole, nous pouvons distinguer des approches différentes. Ces approches peuvent aller de la régression d'un score d'intelligibilité à partir d'un taux d'erreur mot obtenu par un système de reconnaissance automatique de la parole [Christensen 2012], à l'extraction de paramètres pertinents d'une parole pathologique, en utilisant

des technologies de traitement automatique de la parole [Quintas 2020]. Étant donné que les approches basées sur le taux d'erreur mot sont moins performantes sur les patients sévères, et que les approches basées sur le traitement automatique de la parole sont normalement plus difficiles à interpréter, dans [Quintas 2022b, Quintas 2023], nous proposons une méthode automatique pour prédire l'intelligibilité de la parole fondée sur le score individuel de plusieurs pseudo-mots énoncés par un locuteur. Le score final est ainsi calculé en fonction des scores individuels des différents mots prononcés par chaque patient. La méthodologie proposée pour faire la prédiction automatique de l'intelligibilité, repose sur l'utilisation d'un transformer avec un système d'attention [Vaswani 2017b]. Nous pouvons diviser le système en 3 parties distinctes. La première correspond à la **l'extraction des paramètres** : à partir des fichiers audios, nous calculons sur chaque fenêtre 40 filtres (filterbanks). Chaque mot est associé à son score respectif de décodage acoustico-phonétique perceptuel. Ces scores sont utilisés comme les scores perceptifs de référence. La description de l'obtention de ces scores est décrite plus en détail dans la sous-section 3.2.2. La deuxième partie est la **modélisation** : nous utilisons un transformer afin d'obtenir les scores automatiques au niveau de chaque pseudo-mot. Ce type de modèle, proposé par [Vaswani 2017a] et adapté à la reconnaissance vocale par [Dong 2018], suit une architecture encodeur-décodeur. Notre proposition pour le transformer utilise un encodeur récurrent bidirectionnel avec des GRU (Gated Recurrent Units) [Cho 2014]. L'encodeur possède 3 couches récurrentes avec une taille d'entrée de 40 (dimension des filterbanks) et une dimension cachée de 100. La sortie de l'encodeur est suivie d'un mécanisme d'attention. Ce mécanisme permet au système de se concentrer davantage sur des parties particulières du fichier d'entrée, tout en ignorant les parties moins pertinentes. Avec cela, nous espérons que le système apprendra automatiquement des interdépendances intéressantes entre des phonèmes consécutifs et qu'il trouvera le lien entre les erreurs de prononciation et le score DAP de chaque mot. Après le mécanisme d'attention, le vecteur de longueur fixe passe à travers un ensemble de 3 couches entièrement connectées, de dimension $[100 * 100]$ avec des ReLUs (Rectified Linear Units) [Nair 2010] comme fonctions d'activation. Enfin, nous utilisons une couche de Global Max Pooling pour obtenir le score individuel pour chaque mot. L'entraînement de ce système a été effectué grâce à une validation croisée à 10 blocs. À chaque bloc, 113 locuteurs (patients et contrôles) sont utilisés pour l'entraînement, et 13 locuteurs pour l'évaluation. Finalement, la troisième partie est une régression du score général pour déterminer **l'intelligibilité au niveau mots** de chaque locuteur en fonction des scores individuels d'intelligibilité de chaque mot. Ici, nous avons utilisé la moyenne des scores individuels des 52 pseudo-mots de chaque locuteur. La figure 3.14 illustre la chaîne de traitement de notre système.

Évaluation du Système

La corrélation de Spearman (ρ) ainsi que l'erreur quadratique moyenne ($RMSE$, Root Mean Squared Error) nous ont permis d'évaluer notre système. La cible étant les scores perceptuels de DAP mentionnés précédemment en sous-section 3.2.2.

Nos résultats sont présentés dans le tableau 3.5 et illustrés sur la figure 3.15. Ils sont comparés avec une autre approche, provenant de la transcription automatique des pseudo-mots [Fredouille 2019, Ghio 2018]. Celle-ci a obtenu les scores de DAP en utilisant un algorithme Wagner-Fischer entre les scores de référence et les scores obtenus de la transcription automatique. En utilisant les mêmes locuteurs et les mêmes pseudo-mots, nous obtenons de meilleurs résultats : sept points de plus de corrélation et la mesure d'erreur (RMSE) est réduite de près de la moitié.

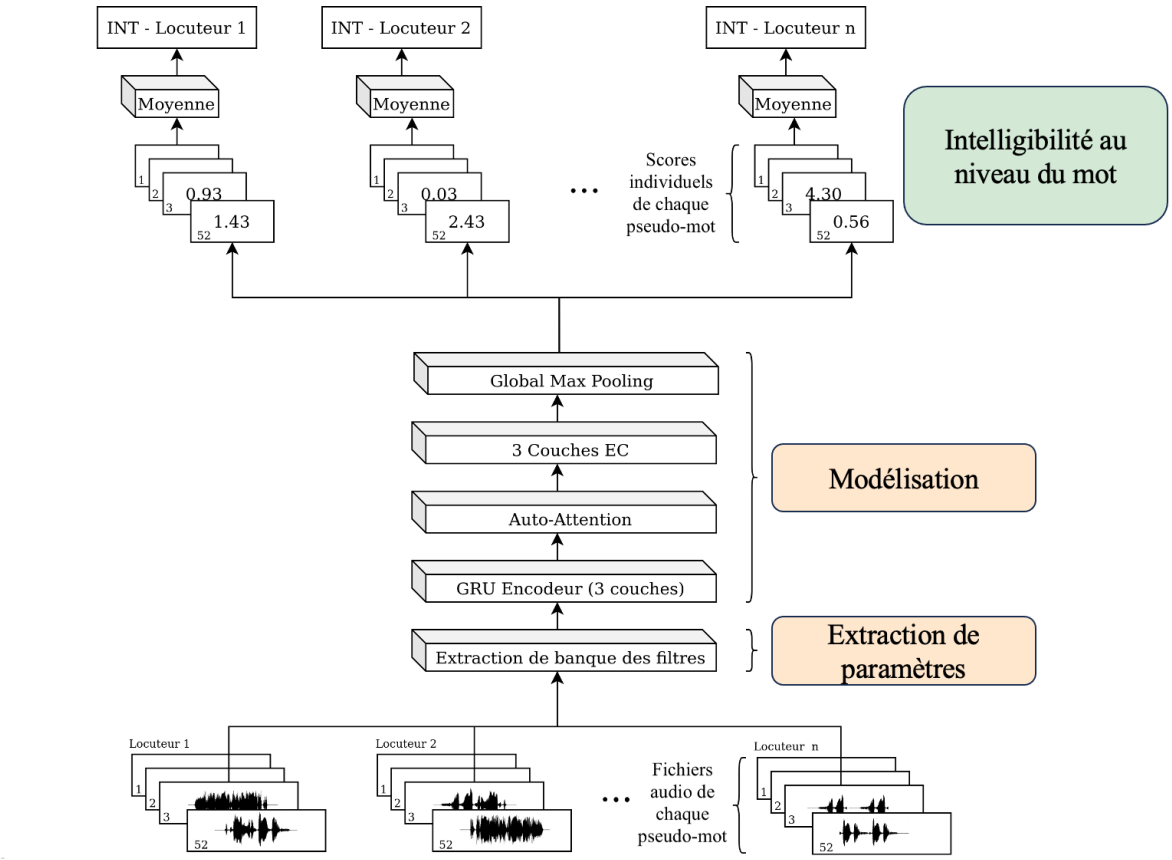

FIGURE 3.14 – Aperçu général du système proposé. EC signifie entièrement connectés. INT signifie le score d'intelligibilité d'un locuteur donné (figure tirée de [Quintas 2022b])

TABLEAU 3.5 – Comparaison entre les résultats de référence et les résultats obtenus avec notre approche

	ρ	$RMSE$
Wagner-Fischer (baseline)	0,80	0,792
Transformer avec auto-attention	0,87	0,370

Réduction de la quantité des pseudo-mots

En regardant de plus près la composition des pseudo-mots utilisés, une idée a émergé. En effet, comme le montre l'encadré 3.5.1, chaque ensemble de 52 mots comporte un sous-ensemble de 16 mots avec double consonne au début, 16 mots avec double consonne au milieu et au moins 26 mots sans double consonne. Les mots peuvent avoir des doubles consonnes à la fois au début et au milieu. De plus, dans un contexte clinique, l'enregistrement de 52 mots est chronophage. Il devient alors pertinent de se demander si l'on peut améliorer les résultats en utilisant moins de mots à la manière de l'article [Marczyk 2020] où les auteurs ont prouvé que nous pouvions réduire la liste aux seuls 16 pseudo-mots ayant une double consonne pour le score DAP des évaluateurs humains. Comment vont se comporter les résultats sur une liste plus restreinte de 16 mots en utilisant la prédiction automatique ?

Le score final de chaque patient est maintenant le score des sous-ensembles des mots de la liste réduite

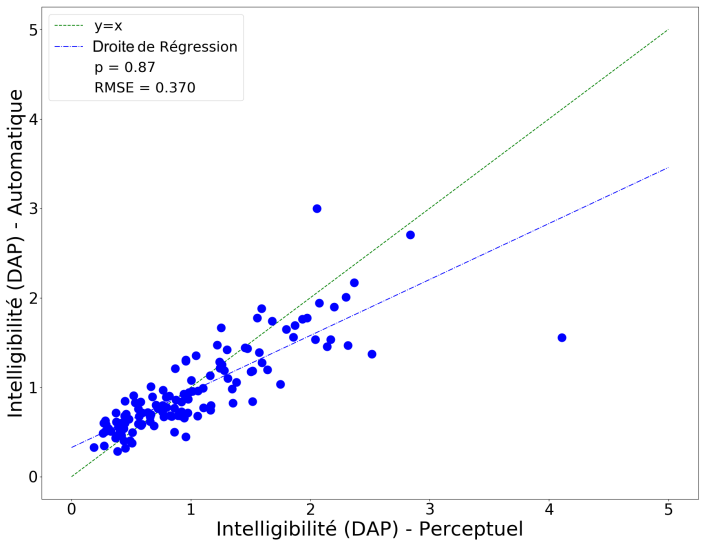

FIGURE 3.15 – Graphique des résultats obtenus, pour la prédiction automatique de l'intelligibilité, tiré de [Quintas 2022b]

Encadré 3.5.1 – Exemple d'un ensemble de 52 pseudo-mots. Le bleu (resp. le violet) correspond aux doubles consonnes en début (resp. en milieu) de pseudo-mots

banfou	**bl**eja	bou**ct**i	**br**im**pl**i	chessant	choniou
clifant	cogu	**cr**impin	daillu	dinrant	**dr**edi
fan**rs**i	**fl**in**rp**u	fouma	**fr**avi	gabi	**gl**unou	**gr**o**rv**o
guchin	joutu	juro	lanvin	le**rd**a	messo	mouco
nianlo	niejo	no**ks**a	nouillou	pa**st**u	pidant
ploniou	**pr**ipin	**ps**ila	quiga	rinta	ru**rn**u
san**vr**in	**sc**una	souquin	**sp**a**cl**ant	**st**icho	tan**gr**i
tou**gz**u	**tr**a**dr**ou	vi**rj**ant	vumou	yainzi
ya**lt**in	zebou	zouzant		

tandis que son score de référence est toujours la moyenne des 52 scores perceptifs. Les résultats sur la liste réduite sont présentés dans le tableau 3.6.

Au niveau de la corrélation et de l'erreur pour les ensembles avec 16 mots, aucun changement majeur n'est observable par rapport aux résultats obtenus avec 52 mots. Cet aspect corrobore le résultat trouvé dans [Marczyk 2020], qu'il est possible d'obtenir une prédiction fiable en utilisant un ensemble de mots significativement plus petit, ici, en utilisant des mesures automatiques.

D'autres résultats sur des ensembles de pseudo-mots se trouvent dans ce tableau. Nous observons par exemple que dans les sous-ensembles de 10 pseudo-mots, il y a un plus grand écart, au niveau de la corrélation et de l'erreur, entre les mots avec et sans double consonne. L'ensemble des 10 pseudo-mots sans double consonne étant le moins bon. Cela corrobore le fait que les mots avec double consonne sont plus pertinents dans l'obtention d'une mesure d'intelligibilité automatique. Ceci est aussi illustré avec les résultats de l'ensemble final de 5 pseudo-mots avec les deux occurrences des doubles consonnes, qui n'ont pas montré de différence importante au niveau de l'erreur par rapport à, par exemple, l'ensemble de 26 pseudo-mots sans double consonne.

TABLEAU 3.6 – Comparaison entre les scores précédemment obtenus sur la liste de pseudos-mots complète et ceux des listes réduites. L'acronyme *d.c.* signifie double consonne

Modèle	**Nombre de pseudo-mots utilisés**	ρ	$RMSE$
Wagner-Fischer (baseline)	52 (total)	0,80	0,792
Transformer avec auto-attention	52 (total)	0,87	0,370
	16 avec *d.c.* au début	**0,85**	**0,370**
	16 avec *d.c.* au milieu	**0,85**	**0,375**
	26 sans *d.c.*	0,84	0,398
	10 avec *d.c.* au début	0,83	0,393
	10 avec *d.c.* au milieu	0,82	0,399
	10 sans *d.c.*	0,76	0,471
	5 avec *d.c.* au début et milieu	**0,79**	**0,413**

Les résultats de ce travail ont montré qu'il est possible d'obtenir une forte corrélation entre une évaluation perceptive et un score automatique obtenu par un modèle transformer avec attention, ceci au niveau du mot. De plus, même en conservant peu de pseudo-mots qui ont un contenu phonétique riche et incitent à la co-articulation de consonnes, les résultats obtenus montrent qu'il est possible d'écourter la liste tout en continuant à obtenir de bonnes prédictions. Cette réduction de la liste à faire prononcer au patient est très pertinente dans le contexte clinique, afin de concevoir des batteries d'examens plus courtes pour les patients.

3.5.3 Modélisation au niveau du phonème

Le troisième niveau de granularité étudié dans [Quintas 2022a] est le niveau du phonème. Ceux-ci, en tant qu'unités minimales distinctives d'un son peuvent être de bons indices pour baser un score d'intelligibilité. De même que dans le système précédent, c'est la tâche de lecture de pseudo-mots (DAP) qui a été utilisée ici, avec le même score d'intelligibilité perçu. Parmi les différents phonèmes du français dont plusieurs occurrences se trouvent être prononcées dans cette tâche, les consonnes semblent jouer un rôle important dans le score d'intelligibilité (voir paragraphe 3.5.2). Ceci nous a conduits à concevoir un système utilisant uniquement les consonnes pour prédire l'intelligibilité. Le score de référence est le même que dans le paragraphe 3.5.1 à savoir le score perceptif de la tâche de description d'image. La chaîne de traitement est décrite dans la figure 3.16 .

Après un alignement forcé effectué grâce au Montreal Forced Alignment [McAuliffe 2017], les différentes consonnes du français sont localisées puis 13 coefficients MFCCs sont extraits. Utilisés pour la détection de mauvaises prononciations en L2 dans [Wang 2018], nous utilisons également un réseau siamois pour évaluer les similarités phonétiques. Le système reçoit deux phones en entrée pour les comparer et les phones sont considérés similaires en fonction d'un seuil (voir figure 3.17). La thèse de Sebastião Quintas [Quintas 2022a] donne plus de détail sur l'architecture utilisée. Chaque consonne émise par un locuteur est donc comparée aux consonnes de l'apprentissage et un score de similarité phonétique est calculé, il correspond au nombre de phonèmes similaires divisés par le nombre total d'occurences du même phone. L'intelligibilité globale est la moyenne des scores de similarité phonétiques de chaque phone p :

$$I(S_p) = \frac{\sum_{n=1}^{16} \frac{Sim_p(n)}{Tot_p(n)}}{16} * 10$$

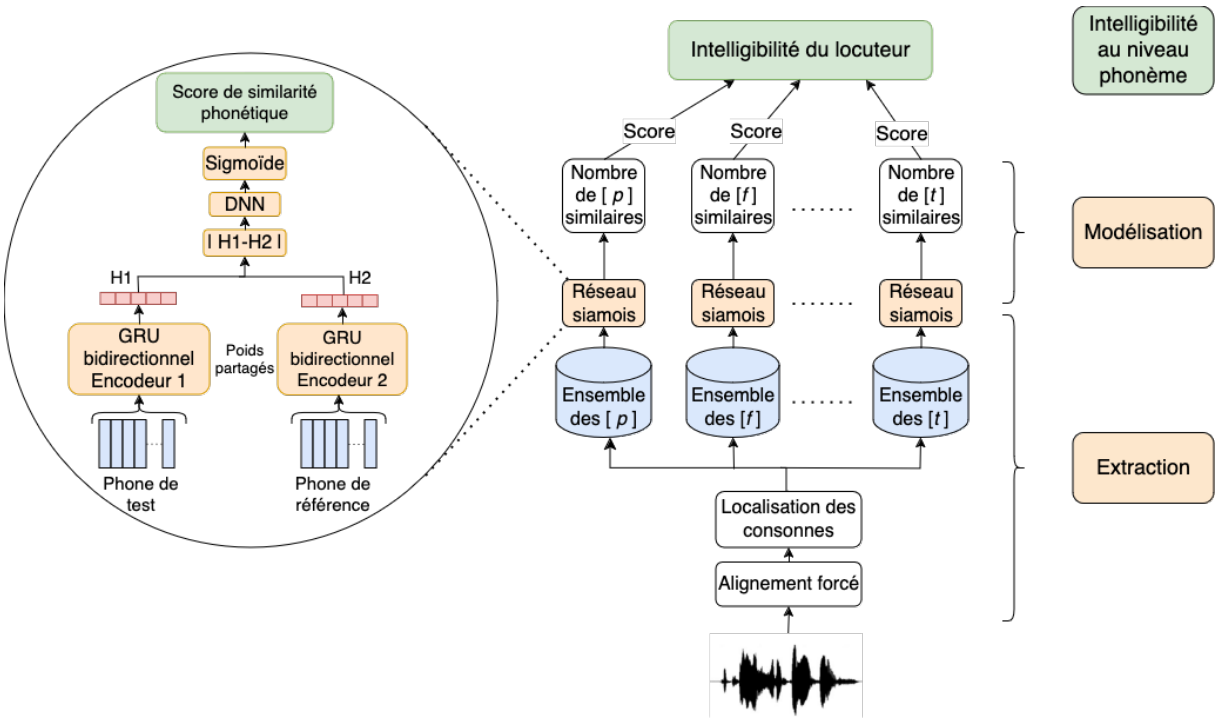

FIGURE 3.16 – Système employé pour prédire l'intelligibilité au niveau phonème

Ici les phonèmes étudiés sont les 16 consonnes du Français.

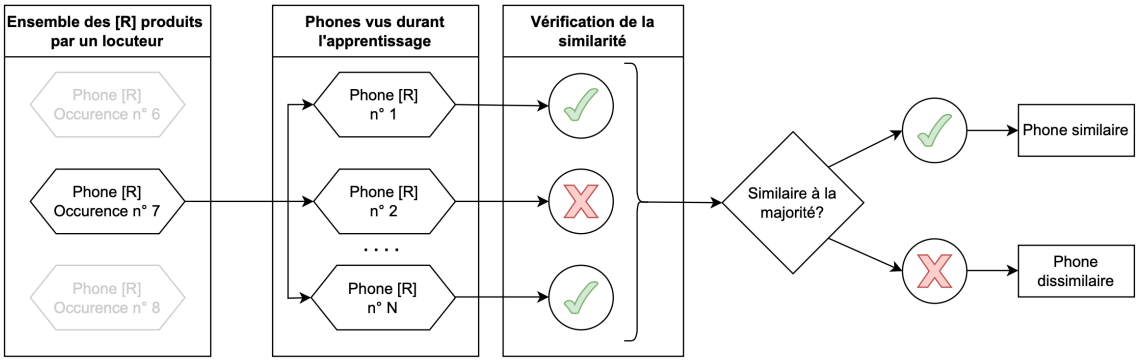

FIGURE 3.17 – Exemple de calcul de la similarité phonétique sur le phone [R]

Comme précédemment, le coefficient de corrélation de Spearman et le Root Mean Square ont été utilisés comme métrique d'évaluation. La figure 3.18 permet de visualiser les résultats de prédiction par locuteur. Un taux de corrélation satisfaisant est obtenu ($p = 0{,}82$), suggérant que l'intelligibilité de la parole peut être obtenue en utilisant la similarité des productions de consonnes entre deux locuteurs.

Dans cette partie sur la modélisation de la parole pathologique, nous avons vu qu'il est possible en s'appuyant sur trois niveaux de granularité que sont la phrase, le mot et le phonème de prédire avec une corrélation élevée l'intelligibilité de la parole d'un patient. De plus, il est à noter que même si au niveau

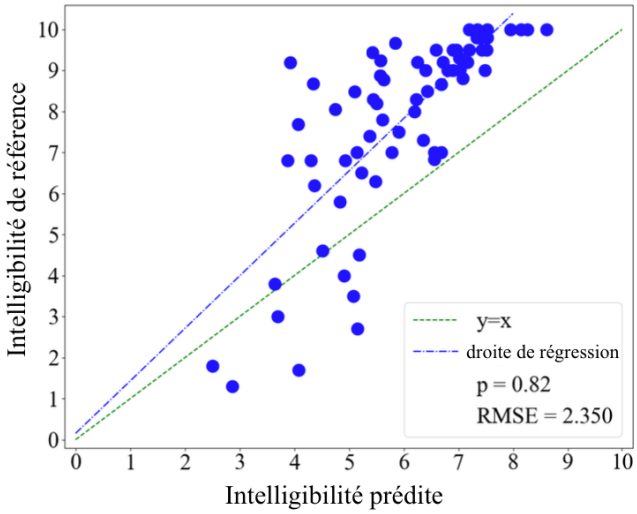

FIGURE 3.18 – Prédiction de l'intelligibilité au niveau phonème de notre système par rapport à l'intelligibilité perçue traduite de [Quintas 2022c]

phonème, le RMS est significativement plus important que pour les autres approches, ce système a été développé sans s'appuyer sur le score d'intelligibilité perçu par des experts en tant que référence. Cet aspect souligne que l'on peut trouver de manière automatique et en s'appuyant sur une modélisation de différence de production de phonèmes, un moyen objectif et explicable de prédire l'intelligibilité.

3.6 Conclusion

Ce chapitre a décrit plusieurs aspects de la parole pathologique en suivant quelques-uns des travaux que j'ai pu mener. En collaboration avec des cliniciens, il a été important de s'attarder sur une constitution de corpus écologique et adaptée aux réalités rencontrées en milieu hospitalier et également de faire un état des lieux sur les outils d'évaluation déjà disponibles en clinique pour évaluer les manques et les besoins. La partie traitement automatique prend alors tout son sens pour venir appuyer les scientifiques et dans plusieurs études telles que celles sur la parole aphasique, la maladie de Parkinson, et les cancers ORL, nous avons vu que les sorties de ce traitement peuvent s'avérer d'une aide précieuse pour de l'aide au diagnostic ou encore de la remédiation. Il s'agira alors de quitter le monde de la recherche pour voir comment déployer ces techniques chez les professionnels de santé.

CHAPITRE 4

PROJET DE RECHERCHE

Sommaire

4.1 Bilan

Tout au long de ce document, j'ai voulu retracer quelques jalons de mes travaux de recherche en mettant en regard les dernières avancées technologiques sur le traitement automatique de la parole standard (ou saine), puis la parole atypique et enfin la parole pathologique, étant le fil rouge de ce manuscrit.

La préparation des données, avec le recueil de corpus et la terminologie à adopter pour la phase d'annotation sont autant d'étapes qu'il faut penser avec soin pour s'assurer de la qualité des futurs résultats (voir chapitre 1). Nous avons vu également que connaître et anticiper les altérations dues à l'environnement dans lequel est produit l'enregistrement de parole (avec l'exemple de la réverbération) permet une bonne prédiction du WER et éventuellement d'aiguiller vers le système de reconnaissance le plus performant et efficace. De plus, le système peut lui-même s'auto-diagnostiquer au moyen de mesures de confiance.

Dans le chapitre suivant (le chapitre 2), l'objectif était de s'intéresser à la parole standard. En effet, dans le cadre du traitement automatique de la parole, la parole standard est la parole la plus souvent contenue dans les corpus d'apprentissage car c'est celle sur laquelle les systèmes les plus performants ont été appris. Dans le contexte de la pathologie, la parole saine est même déjà un concept sur lequel réfléchir, car en fonction des études, parole saine ne veut pas dire parole étalon. L'exemple est pris sur le terme intelligibilité qui est déjà variable sur la parole saine et ne veut pas dire que toutes les personnes saines (ou contrôles) ont une performance à 100%. En parole standard, plusieurs critères peuvent aider à améliorer

les systèmes de TAP. Pour augmenter les données d'un corpus d'apprentissage, nous pouvons utiliser les mesures de confiance vues au premier chapitre pour filtrer de la parole non annotée et prendre celle qui est relativement proche du premier corpus. Ceci va permettre d'ajouter plus d'exemples et de légères variabilités pour augmenter les données du corpus d'apprentissage pour améliorer les performances en WER du système. Les mesures de confiance peuvent également permettre une combinaison de systèmes de reconnaissance de la parole où un premier système propose ses meilleures hypothèses et leurs scores de confiance en entrée d'un deuxième système pour réévaluer dynamiquement les probabilités du modèle de langage de ce dernier. Pour aller plus loin, nous avons abordé dans ce chapitre comment la parole typique peut être variable avec comme exemple la parole d'apprenants japonophones du français. Les mesures automatiques que l'on peut fournir pour évaluer l'écart de prononciation par rapport à un locuteur natif fournissent une aide précieuse à l'apprentissage d'une langue seconde. Ce chapitre montre donc la richesse des défis scientifiques liés à la parole typique dont les variabilités, une fois prises en compte, se révèlent extrêmement utiles pour gagner en performances et en pertinence sur de nombreuses applications.

Un des principaux défis auxquels les praticiens sont actuellement confrontés est l'évaluation des troubles de la parole et de la voix, que ce soit lors du diagnostic initial, ou pour suivre l'évolution de la maladie chez le patient, ou encore pour évaluer l'efficacité des interventions thérapeutiques. L'évaluation perceptive par le praticien est encore à ce jour la méthode la plus couramment utilisée en pratique clinique. Cependant cette méthode est souvent critiquée pour son manque d'objectivité, notamment due au manque de reproductibilité inter et intra-annotateurs. Les progrès réalisés ces dernières années dans le domaine du traitement automatique de la parole offrent une solution potentielle pour apporter une évaluation plus objective et moins coûteuse. Les deux freins principaux sont le manque de corpus de parole pathologique et la nécessité de trouver de nouveaux outils fiables et utiles pour les thérapeutes pour rendre l'évaluation plus objective et reproductible. Le chapitre 3 décrit quelques travaux initiés dans ces voies avec les constitutions de différents corpus de voix pathologiques puis la réflexion menée autour des outils déjà disponibles en clinique pour identifier les besoins. Plusieurs analyses ont ensuite été investiguées. Nous avons vu le cas d'aphasiques dont la parole a été étudiée grâce au nombre de pauses remplies ou non qu'ils produisent. Nous avons également réemployé un algorithme de Goodness of Pronunciation pour analyser la prononciation de paralysés faciaux ou encore utilisés des caractéristiques acoustico-articulatoires pour différencier deux types de syndromes parkinsoniens. La modélisation quant à elle nous a permis de trouver un système capable de déterminer l'intelligibilité d'un énoncé de patients atteints de cancer ORL à trois niveaux de granularité, le phonème, le mot, la phrase.

4.2 Perspectives

En restant dans ces thématiques, nous pouvons continuer à approfondir nos connaissances sur la parole et améliorer les technologies d'assistance et les applications pratiques d'appui aux thérapeutes. Je propose en ce sens plusieurs axes de recherche en ce qui concerne les deux déclinaisons que sont la parole saine et la parole pathologique tout en souhaitant également m'intéresser à la parole atypique.

4.2.1 Sur la parole saine

Pour ce type de parole, plusieurs pistes sont à explorer ou à continuer d'investiguer.

Tout d'abord, pour mieux prendre en compte toutes les formes de discours qui dévient des normes

linguistiques dominantes, qu'il s'agisse d'accents régionaux, de locuteurs non natifs ou encore de personnes avec des particularités phonétiques, l'enrichissement des corpus linguistiques devient essentiel. Il s'agit de collecter et d'intégrer des données de parole plus diversifiées et représentatives dans ces corpus, permettant ainsi une meilleure modélisation et reconnaissance des variations linguistiques. Des méthodes pour enregistrer ces données et décloisonner la collecte sont encore à investiguer. L'initiative « Ecouter Parler » [1] est très intéressante dans le sens où elle propose que la cabine d'enregistrement vienne à la rencontre des populations grâce à un laboratoire mobile dans un camion. Les participants sont ainsi invités à « contribuer au Portrait sonore de la France ». Je souhaiterais effectivement me consacrer plus à ce genre de campagne pour réfléchir à démocratiser « le don de sa propre voix » et aider à faire connaître les besoins des chercheurs en parole en termes de données.

Un des projets sur lequel j'ai également commencé à réfléchir est une application smartphone pour aider les personnes âgées en situation d'isolement et ainsi, améliorer leur cadre de vie au niveau social. Pour rendre les technologies de traitement du langage plus accessibles et utiles pour les populations sous-représentées ou marginalisées comme les personnes âgées, il est crucial de développer des interfaces utilisateur inclusives et des systèmes qui s'adaptent aux spécificités de chaque utilisateur, plutôt que de forcer ces derniers à se conformer à un modèle standardisé. Les techniques d'adaptation au locuteur permettant la personnalisation des systèmes ont été validées comme celle utilisée dans [Gu 2023]. L'un des principaux défis de l'adaptation du locuteur est le manque de données annotées sur le locuteur cible. Une idée est de se servir des énoncés de l'ensemble de d'apprentissage qui ont des caractéristiques vocales similaires à ceux du locuteur cible pour augmenter les données du locuteur cible dans le processus d'adaptation. Des techniques d'embeddings comme les *x-vecteurs* [Snyder 2018a] peuvent alors être utilisées pour représenter le locuteur cible et obtenir une augmentation des performances d'un SRAP pour les énoncés de ce locuteur [Gu 2023]. Valider ces techniques pour faciliter l'accessibilité des personnes âgées aux techniques de reconnaissance vocale et aux assistants vocaux domestiques me semble une nécessité à investiguer dans les prochaines années. Ces personnes sont souvent isolées et n'ont pas été habituées à interagir avec un ordinateur. La parole semble être une piste très intéressante pour les aider à communiquer avec leurs proches et surtout améliorer leur qualité de vie. Accéder aux soins peut être également grandement facilité avec ces techniques.

Face aux variations de la parole, les techniques d'apprentissage par renforcement (APR) montrent également leur utilité [Chen 2022]. Ces modèles sont capables de s'ajuster dynamiquement aux différentes formes de parole, améliorant ainsi leur performance sur des données rares ou variées. En utilisant des modèles pré-entraînés sur des données abondantes et en les adaptant à de nouvelles données plus spécifiques et moins fréquentes, on pourrait obtenir des systèmes plus robustes et inclusifs. Dans le domaine de la RAP, l'APR a été principalement proposé pour s'attaquer aux divergences entre les phases d'apprentissage et de test. Ceci correspond bien au cas où nous aurions un corpus d'apprentissage figé avec de la parole « typique » et que le corpus de test serait lui bruité, ou avec des locuteurs d'accents régionaux... La détérioration des performances vient de deux choses :

1. L'utilisation conventionnelle du critère d'entropie croisée maximise la log-vraisemblance pendant l'entraînement, alors que la performance est calculée par le WER, et non par la log-vraisemblance ;
2. Au cours de la phase d'inférence, le modèle n'est pas confronté à ses propres prédictions.

La méthode APR résout ces deux points en comblant le fossé entre les phases d'apprentissage et de test.

1. https://ecouter-parler.fr/

Par exemple, dans [Chen 2022], les auteurs ont introduit une méthode d'adaptation basée sur l'APR pour une tâche parole à parole, appelée système autocritique (SCST). Cette méthode peut être vue comme un modèle de décision séquentielle, représenté à la figure 4.1. L'ensemble du réseau neuronal encodeur-décodeur est traité comme un agent. À chaque pas de temps t, l'état actuel est formé en concaténant la caractéristique acoustique x_t et la prédiction précédente Y_{t-1}. Le token de sortie sert d'action et met à jour la séquence d'hypothèses générée. Le SCST associe la fonction de perte d'apprentissage et le WER à l'aide d'une fonction de récompense liée au WER. Il calcule la récompense r_t à chaque étape de génération d'un token en en le comparant à la séquence de vérité de terrain Y^*.

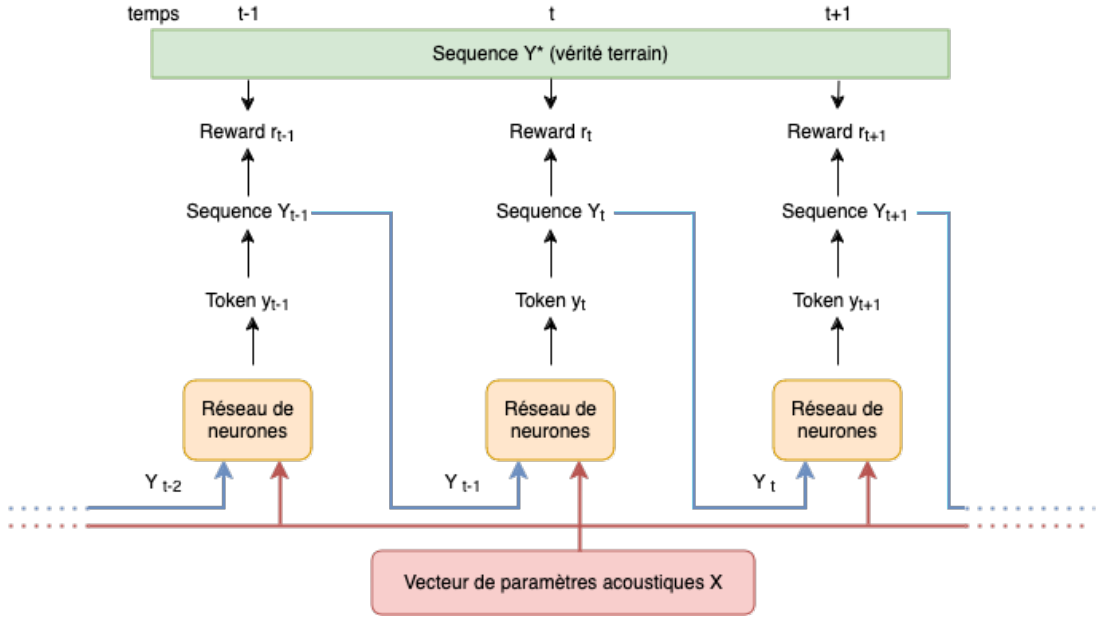

FIGURE 4.1 – Le modèle de décision séquentielle de RAP traduit de [Chen 2022]

Cette méthode a été utilisée avec un corpus de parole propre qui doit être performant sur des séquences de test de parole bruitée. Cette nouvelle méthode reste à valider sur des corpus de test où ce n'est pas l'environnement de production de parole qui provoque une différence avec le corpus d'apprentissage mais lorsque le système est confronté à un énoncé de parole atypique comme de la parole d'enfants ou encore pour créer des systèmes pour aider les apprenants de langue étrangère. Ici, les techniques d'apprentissage profond pourraient modéliser la parole des apprenants dans l'objectif de les aider à atteindre la cible phonétique voulue. Par exemple, le STAP pourrait rejouer leur enregistrement et en leur indiquer précisément quel aspect phonologique ils peuvent travailler. Ceci serait un outil pédagogique très précieux pour la pratique d'une langue et permettrait à l'apprenant d'être plus autonome vis-à-vis du professeur [Detey 2024].

Une autre sorte de déviance observée dans la figure 4.2 est celle de l'émotion convoyée par la parole. Par exemple, dans les situations extrêmes comme les centres d'appel d'urgence ou encore les cabines de pilotage d'avion, plusieurs défis scientifiques sont à relever pour déterminer le stress ou la fatigue dans la voix des pilotes ou encore des opérateurs et des appelants des centres d'appel d'urgence. En vol, le pilote est amené à prendre rapidement les bonnes décisions. Cette lucidité peut être mise à mal lors notamment de vol long qui suscite de la fatigue chez le pilote. Être à même de déclencher des alertes

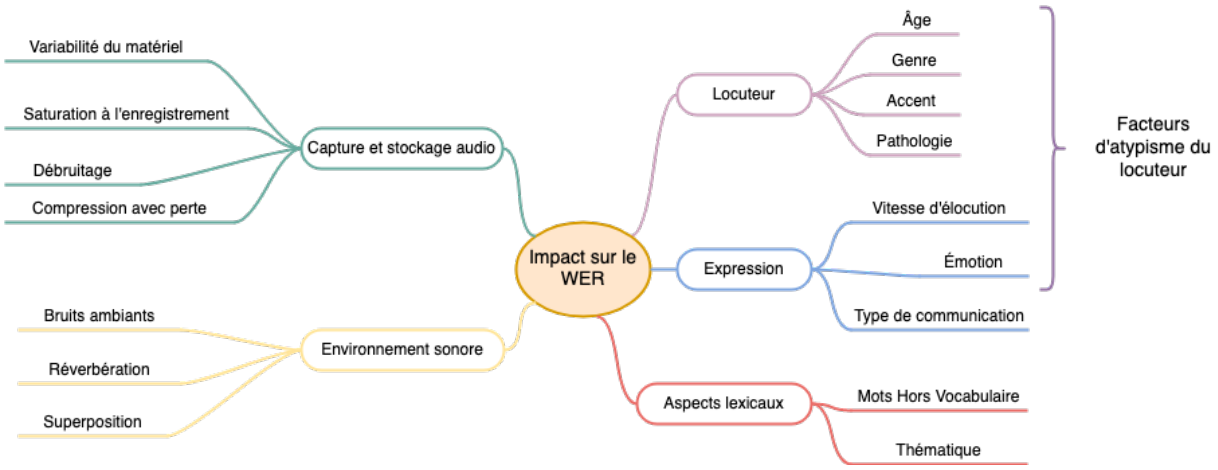

FIGURE 4.2 – Carte heuristique sur les sources de variabilité qui impactent le WER (Word Error Rate ou taux d'erreur mots) des SRAP (systèmes de reconnaissance automatiques de la parole), adaptée de [Ferreira 2021]

grâce à la voix est toujours un défi en cours [Ravi 2019]. Le projet Le Petit Camion qui vient d'être lauréat d'un PRCE de l'ANR évolue justement autour de cette thématique pour déployer un outil d'aide aux opérateurs lors d'appel d'urgence au SDIS (Service Départemental d'Incendie et de Secours) à l'aide de méthodes de traitement automatique de la parole afin de les aider à éviter un oubli crucial dans leur compréhension de l'événement critique en cours. Les effets à l'origine de telles erreurs relèvent des biais cognitifs, dits « de tunnelisation » [Wolff 2018]. Ce projet vise à déterminer dans quelle mesure les technologies modernes de traitement automatique du langage peuvent aider à sécuriser les opérateurs des centres d'appels d'urgence en attirant l'attention de l'opérateur sur de potentielles erreurs ou faiblesses tout en laissant celui-ci seul juge de la situation et seul preneur de décision. La partie du projet où je vais intervenir est celle de la caractérisation de l'environnement et de la parole de l'appelant. L'hypothèse est ici que le degré d'attention de l'opérateur est corrélé avec le niveau de bruit dans le signal audio. Des estimations classiques du rapport signal à bruit et de coefficients de réverbération [Lavechin 2023], ou encore la détection de sons tels que les cris et les pleurs [Gemmeke 2017] ainsi que l'évaluation par l'opérateur lui-même de son degré d'investissement dans la conversation va nous permettre d'explorer ceci. En parallèle, nous étudierons l'impact de l'état émotionnel de l'appelant sur sa parole au travers de la détection de disfluences et d'hésitations [Dutrey 2014], ou de l'estimation de son rythme [Vaysse 2023]. Nous pourrons aussi nous appuyer sur les travaux de [Quintas 2022a] pour mesurer l'intelligibilité des informations fournies par le ou les appelants. Cette tâche intitulée « Informations paralinguistiques et environnementales » est décrite dans la figure 4.3

Ainsi, en enrichissant les corpus linguistiques, en concevant des modèles adaptatifs, et en développant des interfaces inclusives, les technologies de traitement de la parole pourront mieux s'adapter à la diversité des locuteurs et plusieurs de mes projets investigueront sur ces challenges. Mes projets de recherche vont également s'attarder sur le côté parole pathologique, ce qui est décrit dans la section suivante.

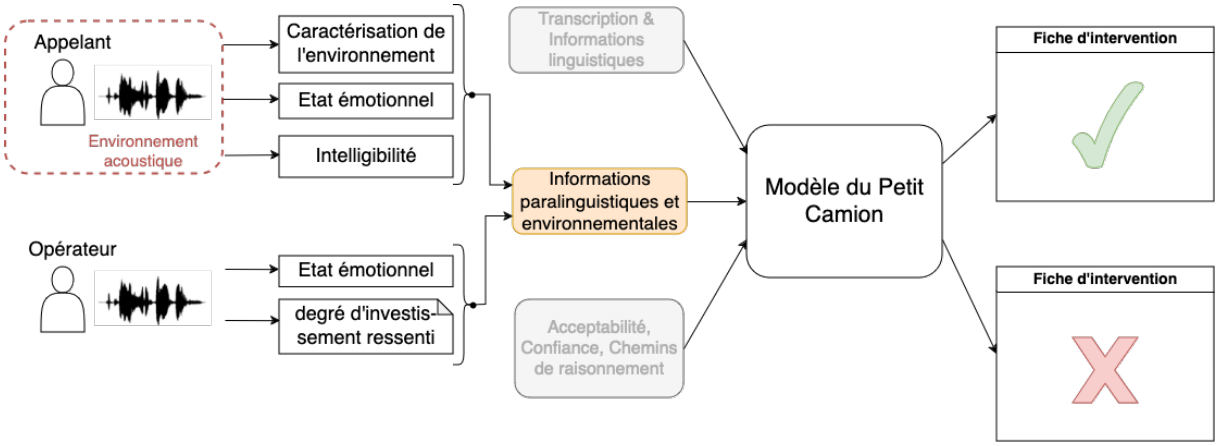

FIGURE 4.3 – Schéma de la tâche « Informations paralinguistiques et environnementales » du projet Le Petit Camion

4.2.2 Sur la parole pathologique

Dans ce domaine, plusieurs champs restent à traiter. Par exemple, les systèmes de traitement automatique peuvent être utilisés pour détecter des anomalies vocales et aider au diagnostic précoce ainsi que dans le suivi des troubles de la parole pathologique. Ils doivent cependant être repensés en collaboration avec les cliniciens pour entrer dans les batteries d'examen en milieu thérapeutique. De même pour les patients, il serait profitable de développer des applications de thérapie vocale capables de leur fournir des retours en temps réel. Sous contrôle d'un thérapeute, ces outils pourraient aider à améliorer la qualité de la rééducation et à motiver les patients grâce à des exercices interactifs et personnalisés. Un premier projet sur lequel se baser est le projet SAMI [Quintas 2024]. Son but est développer une application mobile qui peut être utilisée par les thérapeutes pour évaluer automatiquement l'intelligibilité de la parole des patients. Le principe de l'application est de faire lire à un patient un texte court (moins d'une minute) sur un appareil mobile tel qu'une tablette Android ou iOS. L'enregistrement audio est ensuite traité par un réseau de neurones profonds qui renvoie un score entre 0 (mauvaise intelligibilité) et 10 (intelligibilité parfaite) correspondant à une estimation de l'intelligibilité de la parole. Ce score reproductible peut ensuite être utilisé par les médecins pour se faire une idée de l'impact de la maladie et des traitements (médicaux ou chirurgicaux) sur la qualité de la parole de leurs patients (voir figure 4.4). La fenêtre sur la gauche de la figure montre la tâche que le patient doit lire à haute voix. Ici, il s'agit d'une tâche de lecture du texte « La chèvre de Mr Seguin ». Sur la droite de la figure, nous voyons l'historique des scores obtenus par une même personne sur quelques séances au cours du temps.

L'application a été évaluée sur la production des scores d'intelligibilité en milieu clinique sur 25 patients atteints de cancers ORL, enregistrés en milieu hospitalier [Quintas 2024]. La figure 4.5 est tirée de cet article et montre les résultats obtenus en termes d'intelligibilité sur ce corpus. Le score de référence est obtenu en calculant la moyenne des scores de perceptions d'un panel de 3 juges experts. Une corrélation de 0,818 et une erreur de 1,775 sont obtenues.

Plusieurs pistes restent à poursuivre pour compléter cette application. Tout d'abord, plusieurs mesures autres que le score d'intelligibilité peuvent rentrer dans la conception de SAMI pour la compléter telles que la prosodie, la qualité vocale, ou encore les distorsions phonémiques. Inclure d'autres types de maladies

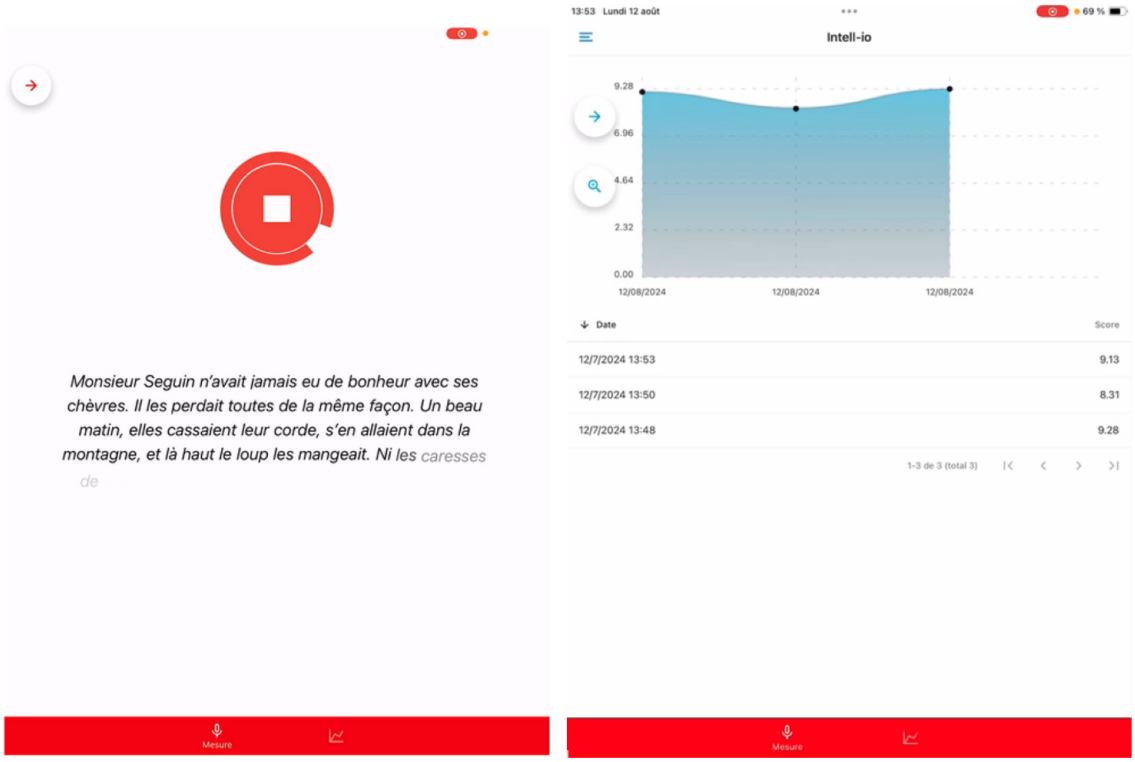

FIGURE 4.4 – Captures d'écrans de l'interface utilisateur de l'application SAMI. À gauche se trouve la partie enregistrement, à droite se trouve l'historique des enregistrements effectués par un même patient ainsi que les scores obtenus

me semble également intéressant tout en continuant à augmenter le premier corpus de patients atteints de cancers ORL pour vérifier le pouvoir de généralisation du modèle. Un gros investissement auprès de la communauté IHM mais pourquoi pas également des concepteurs de jeux vidéos doit également être effectué pour permettre une meilleure prise en main de l'application, faciliter la lecture de la tâche sur l'écran et si l'on veut aller vers un outil d'aide à la remédiation, permettre au patient de rester motivé durant sa thérapie.

Sur un sujet connexe, le projet OLINPICS qui vient d'être lauréat d'un financement ANR PRC, propose de poursuivre les projets C2SI et RUGBI notamment pour continuer d'identifier les unités de parole perturbées induites par des profils phonologiques (apprenants L2) ou phonétiques (pathologies) aux niveaux segmental et prosodique. Une deuxième phase permettra de caractériser les unités de parole perturbées en fonction des styles de parole et des stratégies de compensation individuelle. Ensuite, la phase sur laquelle je vais plus spécifiquement m'attarder est celle de la restauration automatique des unités de parole perturbées. L'objectif est de restaurer la capacité à communiquer des personnes souffrant d'altérations de la voix et de la parole. En effet, perdre la parole pour une personne est socialement préjudiciable, en ce sens qu'elle est un pan entier de sa personnalité. Pour cette raison, l'objectif du projet OLINPIC est de restaurer la parole déficiente en conservant autant que possible la voix du locuteur. La restauration peut être envisagée à des fins prothétiques, thérapeutiques et éducatives ; les méthodes

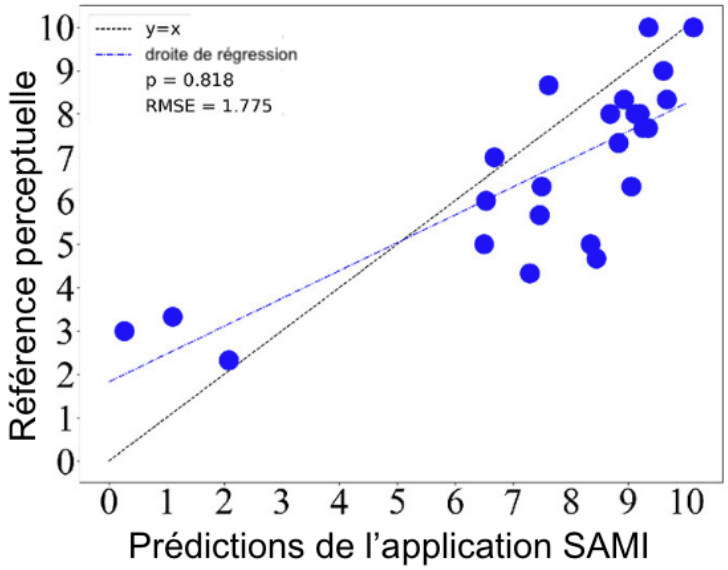

FIGURE 4.5 – Résultats de l'application SAMI en termes de scores d'intelligibilité sur le corpus SAMI tirés de [Quintas 2024]

et les outils pouvant être spécifiquement adaptés à chacun de ces besoins. D'un point de vue automatique, l'un des objectifs ici est de tirer profit des approches basées sur les réseaux de neurones profonds que nous avons spécifiquement conçues pour l'évaluation de l'intelligibilité de la parole altérée dans les projets C2SI et RUGBI [Abderrazek 2022a, Abderrazek 2023, Quintas 2022c]. En effet, les représentations profondes d'une architecture basée sur un réseau de neurones ont montré que les concepts liés aux caractéristiques phonétiques émergent et certaines de ces caractéristiques phonétiques sont dégradées lorsqu'elles sont appliquées à des patients souffrant de différents types de troubles de la parole (cancers HNC, dysarthrie)[Abderrazek 2022b, Quintas 2022c]. Actuellement ces approches sont utilisées pour évaluer la qualité de segments de parole très courts (segments d'une seconde) en termes de qualité d'intelligibilité. Détecter plus finement dans le temps les dégradations de caractéristiques phonétiques permettra de localiser les unités altérées dans le signal de parole pour leur restauration. En ce qui concerne le niveau segmental, une ligne de recherche que nous envisageons de poursuivre consistera à tirer parti des avancées récentes en matière de reconstruction de la parole altérée consacrée à la dysarthrie (DSR - Dysarthria Speech Reconstruction), principalement basée sur la conversion de la voix [Sisman 2020]. Par exemple les modèles encodeurs-décodeurs neuronaux de [Wang 2022] considèrent les différentes composantes de la parole (contenu, prosodie, identité du locuteur) séparément, et permettent de les reconstruire de manière plus explicite. En se tournant vers le domaine de la traduction parole vers parole [Jia 2022] couplée à la préservation de la voix du locuteur, nous pourrons également considérer les segments de parole dégradée à restaurer comme la langue d'origine et la parole normale comme la langue cible.

Finalement, j'espère que d'ici une dizaine d'année, les méthodes automatiques soient largement acceptées et intégrées parmi les outils d'évaluation des thérapeutes et même de médecins généralistes pour diagnostiquer divers troubles. Grâce aux avancées des technologies de traitement de la parole, ces systèmes seront capables d'analyser finement les caractéristiques de la parole d'un patient, de détecter des signes subtils à un stade précoce. Ces outils apporteront un soutien précieux aux professionnels de santé

en offrant des analyses objectives et reproductibles des symptômes en temps réel, complétant ainsi les méthodes traditionnelles d'évaluation. Leur utilisation facilitera également un suivi continu des patients, permettant d'affiner les diagnostics et d'adapter les traitements de façon personnalisée, rapide et efficace.

TABLE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ENCADRÉS

BIBLIOGRAPHIE

[Abderrazek 2022a] Abderrazek S., Fredouille C., Ghio A., Lalain M., Meunier C. et Woisard V., Towards interpreting deep learning models to understand loss of speech intelligibility in speech disorders step 2 : Contribution of the emergence of phonetic traits, dans *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7387–7391, 2022a.

[Abderrazek 2022b] Abderrazek S., Fredouille C., Ghio A., Lalain M., Meunier C. et Woisard V., Validation of the neuro-concept detector framework for the characterization of speech disorders : A comparative study including dysarthria and dysphonia, dans *Interspeech 2022*, pages 3638–3642, 2022b.

[Abderrazek 2023] Abderrazek S., Fredouille C., Ghio A., Lalain M., Meunier C. et Woisard V., Interpreting deep representations of phonetic features via neuro-based concept detector : Application to speech disorders due to head and neck cancer, *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31 :200–214, 2023.

[Ananthapadmanabha 1979] Ananthapadmanabha T. et Yegnanarayana B., Epoch extraction from linear prediction residual for identification of closed glottis interval, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 27(4) :309–319, 1979.

[André-Obrecht 1988] André-Obrecht R., A new statistical approach for the automatic segmentation of continuous speech signals, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 36(1) :29–40, 1988.

[Astésano 2018] Astésano C., Balaguer M., Farinas J., Fredouille C., Ghio A., Gaillard P., Giusti L., Laaridh I., Lalain M., Lepage B. et et al. J. M., Carcinologic speech severity index project : A database of speech disorder productions to assess quality of life related to speech after cancer, *Language Resources and Evaluation Conference*, 2018.

[Auzou 2006] Auzou P. et Rolland-Monnoury V., *BECD : batterie d'évaluation clinique de la dysarthrie*, Ortho édition, 2006, URL https://books.google.fr/books?id=nTR2jgEACAAJ.

[Bazillon 2008] Bazillon T., Estève Y. et Luzzati D., Transcription manuelle vs assistée de la parole préparée et spontanée, dans *JEP 2008*, Avignon, France, 2008, URL https://hal.archives-ouvertes.fr/hal-01450913.

[Benzeghiba 2007] Benzeghiba M., De Mori R., Deroo O., Dupont S., Erbes T., Jouvet D., Fissore L., Laface P., Mertins A., Ris C., Rose R., Tyagi V. et Wellekens C., Automatic speech recognition and

speech variability : A review, *Speech Communication*, 49(10) :763–786, 2007, URL https://www.sciencedirect.com/science/article/pii/S0167639307000404, intrinsic Speech Variations.

[Birko 2015] Birko S., Dove E. S. et Azdemir V., Evaluation of nine consensus indices in delphi foresight research and their dependency on delphi survey characteristics : A simulation study and debate on delphi design and interpretation, *PLOS ONE*, 10(8) :1–14, 08 2015, URL https://doi.org/10.1371/journal.pone.0135162.

[Bond 1994] Bond Z. et Moore T. J., A note on the acoustic-phonetic characteristics of inadvertently clear speech, *Speech Communication*, 14(4) :325–337, 1994, URL https://www.sciencedirect.com/science/article/pii/0167639394900264.

[Bradlow 1996] Bradlow A., Torretta G. et Pisoni D., Intelligibility of normal speech i : Global and fine-grained acoustic-phonetic talker characteristics, *Speech Comm.*, 20 :255–272, 12 1996.

[Chalmers 2019] Chalmers J. et Armour M., *The Delphi Technique*, pages 715–735, Springer Singapore, Singapore, 2019, URL https://doi.org/10.1007/978-981-10-5251-4_99.

[Chen 2022] Chen C., Hu Y., Hou N., Qi X., Zou H. et Chng E., Self-critical sequence training for automatic speech recognition, dans *ICASSP 2022*, pages 3688–3692, 04 2022.

[Cho 2014] Cho K., van Merriënboer B., Bahdanau D. et Bengio Y., On the properties of neural machine translation : Encoder–decoder approaches, dans Wu D., Carpuat M., Carreras X. et Vecchi E. M., rédacteurs, *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*, pages 103–111, Association for Computational Linguistics, Doha, Qatar, Octobre 2014, URL https://aclanthology.org/W14-4012.

[Christensen 2012] Christensen H., Cunningham S., Fox C., Green P. et Hain T., A comparative study of adaptive, automatic recognition of disordered speech, *Proceedings of Interspeech*, 2012.

[Collis 2012] Collis J. et Bloch S., Survey of uk speech and language therapists' assessment and treatment practices for people with progressive dysarthria, *International Journal of Language & Communication Disorders*, 47(6) :725–737, 2012, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1460-6984.2012.00183.x.

[Conway 2015] Conway A. et Walshe M., Management of non-progressive dysarthria : practice patterns of speech and language therapists in the republic of ireland, *International Journal of Language & Communication Disorders*, 50(3) :374–388, 2015, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1460-6984.12143.

[Cox 2002] Cox S. et Dasmahapatra S., High-level approaches to confidence estimation in speech recognition, *IEEE Transactions on Speech and Audio Processing*, 10(7), 2002.

[Darley 1975] Darley F. L., Aronson A. E. et Brown J. R., *Motor speech disorders*, W. B. Saunders and Co., Philadelphia, 1975.

[Deléglise 2005] Deléglise P., Estève Y., Meignier S. et Merlin T., The LIUM speech transcription system : a CMU Sphinx III-based system for french broadcast news, dans *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, Septembre 2005.

[Denman 2019] Denman D., Kim J.-H., Munro N., Speyer R. et Cordier R., Describing language assessments for school-aged children : a delphi study, *International Journal of Speech-Language Pathology*, 21(6) :602–612, Décembre 2019.

[Detey 2024] Detey S., De Fino V. et Fontan L., Morphophonological ambiguities and automatic assessment of spoken l2 lexical forms for pedagogical purposes : a pilot study among japanese learners of french, dans *European Second Language Association 2024*, 07 2024.

[Detey 2005] Detey S., Durand J. et Nespoulous J.-L., Interphonologie et représentations orthographiques. Le cas des catégories /b/ et /v/ chez des apprenants japonais de Français Langue Etrangère., *Revue PAROLE*, 34-35-36 :140–185, 2005, URL https://shs.hal.science/halshs-00274620.

[Dong 2018] Dong L., Xu S. et Xu B., Speech-transformer : A no-recurrence sequence-to-sequence model for speech recognition, dans *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5884–5888, 2018.

[Dumortier 2014] Dumortier B. et Vincent E., Blind rt60 estimation robust across room sizes and source distances, dans *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5187–5191, 2014.

[Dutrey 2014] Dutrey C., *Analyse et détection automatique de disfluences dans la parole spontanée conversationnelle*, Theses, Université Paris Sud - Paris XI, Décembre 2014, URL https://theses.hal.science/tel-01164385.

[Falk 2010] Falk T., Zheng C. et Chan W.-Y., A non-intrusive quality and intelligibility measure of reverberant and dereverberated speech, *Audio, Speech, and Language Processing, IEEE Transactions on*, 18 :1766 – 1774, 10 2010.

[Farinas 2002] Farinas J., *Une modélisation automatique du rythme pour l'identification des langues* , Thèse de doctorat, Université Paul Sabatier, Toulouse, France, novembre 2002.

[Ferreira 2021] Ferreira S., *Prédiction a priori de la qualité de la transcription automatique de la parole par l'analyse de l'environnement sonore*, Thèse de doctorat, École doctorale Mathématiques, Informatique et Télécommunications de Toulouse, 2021.

[Ferreira 2020a] Ferreira S., Farinas J., Pinquier J., Mauclair J. et Rabant S., Analyse de l'effet de la réverbération sur la reconnaissance automatique de la parole, dans Benzitoun C., Braud C., Huber L., Langlois D., Ouni S., Pogodalla S. et Schneider S., rédacteurs, *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition).*, volume 1, pages 235–243, ATALA, Nancy, France, 2020a, URL https://hal.archives-ouvertes.fr/hal-02798542.

[Ferreira 2020b] Ferreira S., Farinas J., Pinquier J., Mauclair J. et Rabant S., Une nouvelle mesure de la réverbération pour prédire les performances a priori de la transcription de la parole, dans Benzitoun C., Braud C., Huber L., Langlois D., Ouni S., Pogodalla S. et Schneider S., rédacteurs, *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition).*, volume 1, pages 226–234, ATALA, Nancy, France, 2020b, URL https://hal.archives-ouvertes.fr/hal-02798541.

[Fredouille 2019] Fredouille C., Ghio A., Laaridh I., Lalain M. et Woisard V., Acoustic-phonetic decoding for speech intelligibility evaluation in the context of head and neck cancers, *International Congress of Phonetic Sciences (ICPhS)*, 2019.

[Gatignol 2004] Gatignol P. et Lamas G., *Paralysies faciales*, Solal Editions, 2004.

[Gemmeke 2017] Gemmeke J. F., Ellis D. P. W., Freedman D., Jansen A., Lawrence W., Moore R. C., Plakal M. et Ritter M., Audio set : An ontology and human-labeled dataset for audio events, dans *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780, 2017.

[Ghannay 2015] Ghannay S., Estève Y. et Camelin N., Word embeddings combination and neural networks for robustness in asr error detection, dans *2015 23rd European Signal Processing Conference (EUSIPCO)*, pages 1671–1675, 2015.

[Ghio 2018] Ghio A., Lalain M., Giusti L., Pouchoulin G., Robert D., Rebourg M., Fredouille C., Laaridh I. et Woisard V., Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique, dans *XXXIIe Journées d'Etudes sur la Parole*, pages 285–293, LPL, ISCA, Aix-en-Provence, France, 2018, URL https://hal.science/hal-01770161.

[Ghio 2021] Ghio A., Pouchoulin G., Viallet F., Giovanni A., Woisard V., Crevier-Buchman L., Hirsch F., Fauth C. et Fredouille C., Du recueil à l'exploitation des corpus de parole " pathologique " : comment accéder à la variation physiopathologique ?, *Corpus*, 22, Janvier 2021, URL https://hal.science/hal-03145102.

[Gillespie 2001] Gillespie B., Malvar H. et Florencio D., Speech dereverberation via maximum-kurtosis subband adaptive filtering, dans *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No.01CH37221)*, volume 6, pages 3701–3704, 2001.

[Goldman 2011] Goldman J.-P., Easyalign : an automatic phonetic alignment tool under praat, dans *Proc. Interspeech 2011*, pages 3233–3236, 2011.

[Gravier 2004] Gravier G., Bonastre J.-F., Galliano S. et Geoffrois E., The ESTER evaluation campaign of rich transcription of french broadcast news, dans *LREC, Language Evaluation and Resources Conference*, Lisbonne, Portugal, Mai 2004.

[Gu 2023] Gu Y., Du Z., Zhang S., Chen Q. et Han J., Personality-aware training based speaker adaptation for end-to-end speech recognition, dans *Interspeech 2023*, pages 1249–1253, 08 2023.

[Guenther 1995] Guenther F. H., Speech sound acquisition, coarticulation, and rate effects in a neural network model of speech production, *Psychological Review*, 102 :594–621, 1995.

[Gurevich 2017] Gurevich N. et Scamihorn S. L., Speech-language pathologists' use of intelligibility measures in adults with dysarthria, *American Journal of Speech-Language Pathology*, 26(3) :873–892, 2017, URL https://pubs.asha.org/doi/abs/10.1044/2017_AJSLP-16-0112.

[Hartman 1995] Hartman F. T. et Baldwin A., Using technology to improve delphi method, *Journal of Computing in Civil Engineering*, 9(4) :244 ?249, 1995.

[Hazan 2017] Hazan V., Speech communication across the lifespan, *Acoustics Today*, 13 :36–43, 04 2017.

[House 1985] House J. et Brackmann D., Facial nerve grading system, *Otolaryngology-Head and Neck Surgery*, 93 :146 – 147, 1985.

[Jacobi 2013] Jacobi I., van Rossum M. A. et van der Molen L., Acoustic analysis of changes in articulation proficiency in patients with advanced head and neck cancer treated with chemoradiotherapy, *The annals of Otology, Rhinology and Laryngology*, 2013.

[Jia 2022] Jia Y., Ramanovich M. T., Remez T. et Pomerantz R., Translatotron 2 : High-quality direct speech-to-speech translation with voice preservation, dans Chaudhuri K., Jegelka S., Song L., Szepesvari C., Niu G. et Sabato S., rédacteurs, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 de *Proceedings of Machine Learning Research*, pages 10120–10134, PMLR, 17–23 Jul 2022, URL https://proceedings.mlr.press/v162/jia22b.html.

[Jousse 2008] Jousse V., Estève Y., Béchet F., Bazillon T. et Linares G., Caractérisation et détection de parole spontanée dans de larges collections de documents audio, dans *JEP*, Avignon, France, Juin 2008, URL https://hal.archives-ouvertes.fr/hal-01321187.

[Kanters 2009] Kanters S., Cucchiarini C. et Strik H., The goodness of pronunciation algorithm : a detailed performance study, dans *SLaTE, ISCA Workshop on Speech and Laguage Technology*, pages 2–5, Wroxall Abbey Estate, Warwickshire, England, 2009.

[Kinoshita 2013] Kinoshita K., Delcroix M., Yoshioka T., Nakatani T., Habets E., Haeb-Umbach R., Leutnant V., Sehr A., Kellermann W., Maas R., Gannot S. et Raj B., The reverb challenge : A common evaluation framework for dereverberation and recognition of reverberant speech, dans *Proceedings of Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 10 2013.

[Kuruvilla-Dugdale 2020] Kuruvilla-Dugdale M., Dietrich M., McKinley J. D. et Deroche C., An exploratory model of speech intelligibility for healthy aging based on phonatory and articulatory measures, *Journal of Communication Disorders*, 87 :105995, 2020, URL https://www.sciencedirect.com/science/article/pii/S0021992420300630.

[Laaridh 2018] Laaridh I., Fredouille C., Ghio A., Lalain M. et Woisard V., Automatic evaluation of speech intelligibility based on i-vectors in the context of head and neck cancers, *Proceedings of Interspeech*, 2018.

[Laaridh 2020] Laaridh I. et Mauclair J., Sur l'utilisation de la reconnaissance automatique de la parole pour l'aide au diagnostic différentiel entre la maladie de Parkinson et l'AMS, dans Benzitoun C., Braud C., Huber L., Langlois D., Ouni S., Pogodalla S. et Schneider S., rédacteurs, *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 1 : Journées d'Études sur la Parole*, volume 1 de *Volume 1 : Journées d'Études sur la Parole*, pages 299–307, ATALA, Nancy, France, 2020, URL https://hal.science/hal-02798552.

[Laborde 2016] Laborde V., Pellegrini T., Fontan L., Mauclair J., Sahraoui H. et Farinas J., Pronunciation assessment of Japanese learners of French with GOP scores and phonetic information, dans *Annual conference Interspeech (INTERSPEECH 2016)*, pages 2686–2690, San Francisco, CA, United States, Septembre 2016, URL https://hal.archives-ouvertes.fr/hal-01474896.

[Lalain 2020] Lalain M., Ghio A., Giusti L., Robert D., Fredouille C. et Woisard V., Design and development of a speech intelligibility test based on pseudowords in french : Why and how ?, *Journal of Speech, Language, and Hearing Research*, 63(7) :2070–2083, 2020.

[Lamel 1993] Lamel L., Gauvain J.-L. et Eskenazi M., Bref, a large vocabulary spoken corpus for french, dans *Eurospeech*, pages 505–508, Genoa, Italy, November 1993.

[Lavechin 2023] Lavechin M., Métais M., Titeux H., Boissonnet A., Copet J., Rivière M., Bergelson E., Cristia A., Dupoux E. et Bredin H., Brouhaha : Multi-task training for voice activity detection,

speech-to-noise ratio, and c50 room acoustics estimation, dans *IEEE Automatic Speech Recognition and Understanding Workshop 2023*, pages 1–7, 12 2023.

[Lecouteux 2007] Lecouteux B., Linarès G., Estève Y. et Mauclair J., System Combination by Driven Decoding, dans *IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07* , Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on, Honolulu, United States, Avril 2007, URL https://hal.archives-ouvertes.fr/hal-01318073.

[Lee 2001] Lee C.-H., Statistical confidence measures and their applications, dans *Proc. of ICSP*, pages 1021–1028, Daejon, Corée du Sud, Aoè»t 2001.

[Lindblom 1990] Lindblom B., Explaining phonetic variation : A sketch of the h&h theory, dans Hardcastle W. J. et Marchal A., rédacteurs, *Speech Production and Speech Modelling*, pages 403–439, Springer Netherlands, Dordrecht, 1990, URL https://doi.org/10.1007/978-94-009-2037-8_16.

[Linstone 2002a] Linstone H. A. et Turoff M., Computers and the future of delphi : Introduction, *The Delphi method : Techniques and applications*, page 483 ?489, 2002a.

[Linstone 2002b] Linstone H. A. et Turoff M., The delphi method : Techniques and applications, *New Jersey Institute of Technology*, 2002b.

[Marczyk 2020] Marczyk A., Ghio A., Lalain M., Rebourg M., Fredouille C. et Woisard V., Have a cake and eat it too : Assessing discrimination performance of an intelligibility index obtained from a reduced sample size, *12th Conference on Language Resources and Evaluation*, 2020.

[Mauclair 2006a] Mauclair J., *Mesures de confiance en traitement automatique de la parole et applications*, Thèse de doctorat, Université du Maine, Décembre 2006a.

[Mauclair 2006b] Mauclair J., Estève Y., Petit-Renaud S. et Deléglise P., Automatic detection of well recognized words in automatic speech transcriptions, dans *LREC, Language Resources and Evaluation*, Genoa, Italy, Mai 2006b.

[McAuliffe 2017] McAuliffe M., Socolof M., Mihuc S., Wagner M. et Sonderegger M., Montreal forced aligner : Trainable text-speech alignment using kaldi, dans *Interspeech*, 2017, URL https://api.semanticscholar.org/CorpusID:12418404.

[McCloy 2015] McCloy D. R., Wright R. A. et Souza P. E., Talker versus dialect effects on speech intelligibility : A symmetrical study, *Language and Speech*, 58(3) :371–386, 2015.

[Metz 1990] Metz D. E., Schiavetti N., Samar V. J. et Sitler R. W., Acoustic dimensions of hearing-impaired speakers' intelligibility : segmental and suprasegmental characteristics., *Journal of speech and hearing research*, 33 3 :476–487, 1990.

[Metze 2000] Metze F., Kemp T., Schaaf T., Schultz T. et Soltau H., Confidence measure based language identification, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Istanbul, Turkey, Juin 2000.

[Meunier 2007] Meunier C., Phonétique acoustique, dans P. A., rédacteur, *Les dysarthries*, pages 164–173, Solal, 2007, URL https://hal.archives-ouvertes.fr/hal-00250272.

[Mikolov 2013] Mikolov T., Sutskever I., Chen K., Corrado G. et Dean J., Distributed representations of words and phrases and their compositionality, dans *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, Curran Associates Inc., Red Hook, NY, USA, 2013.

[Miller 2017] Miller N. et Bloch S., A survey of speech langage therapy provision for people with post-stroke dysarthria in the uk, *International Journal of Language & Communication Disorders*, 52(6) :800–815, 2017, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1460-6984.12316.

[Miller 2019] Miller T., Explanation in artificial intelligence : Insights from the social sciences, *Artificial Intelligence*, 267 :1–38, 2019, URL https://www.sciencedirect.com/science/article/pii/S0004370218305988.

[Nair 2010] Nair V. et Hinton G. E., Rectified linear units improve restricted boltzmann machines, dans *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807 ?814, Omnipress, Madison, WI, USA, 2010.

[Nocaudie 2018] Nocaudie O., Astésano C., Ghio A., Lalain M. et Woisard V., 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 *XXXIIe Journées d'Etudes sur la Parole*, pages 196–204, Aix-en-Provence, France, Juin 2018, URL https://hal.science/hal-01962272.

[Nocera 2002] Nocera P., Linarès G. et Dominique M., Principes et performances du décodeur parole continue Speeral, dans *JEP*, Nancy, France, Juin 2002, URL https://hal.archives-ouvertes.fr/hal-01319843.

[Oomen 2001] Oomen C. E., Postma A. et Kolk H. H. J., Pre-articulary and post-articulary self-monitoring in broca ?s aphasia, *Cortex*, 37(5) :627–641, 2001.

[Pappagari 2020] Pappagari R., Wang T., Villalba J., Chen N. et Dehak N., X-vectors meet emotions : A study on dependencies between emotion and speaker recognition, *Proceedings of ICASSP*, 2020.

[Paul 1992] Paul D. B. et Baker J. M., The design for the Wall Street Journal-based CSR corpus, dans *Speech and Natural Language : Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992, URL https://aclanthology.org/H92-1073.

[Pellegrini 2015] Pellegrini T., Fontan L., Mauclair J., Farinas J., Alazard-Guiu C., Robert M. et Gatignol P., Automatic Assessment of Speech Capability Loss in Disordered Speech, *ACM Transactions on Accessible Computing* , 6(3) :1–14, Mai 2015, URL https://hal.archives-ouvertes.fr/hal-01371812.

[Pellegrini 2014] Pellegrini T., Fontan L., Mauclair J., Farinas J. et Robert M., The goodness of pronunciation algorithm applied to disordered speech, dans *The 15th Annual Conference of the International Speech Communication Association - INTERSPEECH 2014*, pages 1463–1467, International Speech Communication Association (ISCA), Singapore, SG, 2014, URL https://oatao.univ-toulouse.fr/13139/.

[Petrick 2008] Petrick R., Lohde K., Lorenz M. et Hoffmann R., A new feature analysis method for robust asr in reverberant environments based on the harmonic structure of speech, dans *2008 16th European Signal Processing Conference*, pages 1–5, 2008.

[Pommée 2021a] Pommée T., *Les mesures d'intelligibilité : Etat de l'art, considérations pratiques pour l'applicabilité clinique et explorations acoustiques*, Thèse de doctorat, École doctorale Mathématiques, Informatique et Télécommunications de Toulouse, 2021a, URL http://www.theses.fr/2021TOU30141/document, thèse de doctorat dirigée par Pinquier, Julien et Woisard, Virginie Informatique et télécommunications Toulouse 3 2021.

[Pommée 2021b] Pommée T., Balaguer M., Mauclair J., Pinquier J. et Woisard V., Intelligibility and comprehensibility : A delphi consensus study, *International Journal of Language & Communication Disorders*, 57(1) :21–41, 2021b, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/1460-6984.12672.

[Pommée 2022] Pommée T., Balaguer M., Mauclair J., Pinquier J. et Woisard V., Assessment of adult speech disorders : current situation and needs in french-speaking clinical practice, *Logopedics Phoniatrics Vocology*, 47(2) :92–108, 2022, URL https://doi.org/10.1080/14015439.2020.1870245, pMID : 33423572.

[Pommée 2021c] Pommée T., Balaguer M., Pinquier J., Mauclair J., Woisard V. et Speyer R., Relationship between phoneme-level spectral acoustics and speech intelligibility in healthy speech : a systematic review, *Speech, Language and Hearing*, 24(2) :105–132, 2021c, URL https://doi.org/10.1080/2050571X.2021.1913300.

[Pommée 2024] Pommée T., Bouvier L., Pinquier J., Mauclair J., Delvaux V., Fougeron C., Astésano C., Martel-Sauvageau V., Morsomme D., Pinçon P., Lalain M. et Woisard V., Le voyage d'alice : un texte standardisé pour l'évaluation de la parole et de la voix en français., *Glossa*, 138 :6–43, 01 2024.

[Postma 2000] Postma A., Detection of errors during speech production. a review of speech monitoring models, *Cognition*, 77(2) :97–131, 2000.

[Povey 2011] Povey D., Ghoshal A., Boulianne G., Burget L., Glembek O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P. et al., The Kaldi speech recognition toolkit, dans *IEEE 2011 Workshop on automatic speech recognition and understanding*, IEEE Signal Processing Society, 2011.

[Quintas 2022a] Quintas S., *Deep learning approaches to assess speech intelligibility of head and neck cancer*, Theses, Université Paul Sabatier - Toulouse III, Novembre 2022a, URL https://theses.hal.science/tel-04094765.

[Quintas 2022b] Quintas S., Abad A., Mauclair J., Woisard V. et Pinquier J., Utilisation de réseaux de neurones profonds avec attention pour la prédiction de l'intelligibilité de la parole de patients atteints de cancers ORL, dans *Proc. XXXIVe Journées d'Études sur la Parole – JEP 2022*, pages 63–71, 2022b.

[Quintas 2023] Quintas S., Abad A., Mauclair J., Woisard V. et Pinquier J., Towards reducing patient effort for the automatic prediction of speech intelligibility in head and neck cancers, dans *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023.

[Quintas 2020] Quintas S., Mauclair J., Woisard V. et Pinquier J., Automatic prediction of speech intelligibility based on x-vectors in the context of head and neck cancer, *Proceedings of Interspeech, Shanghai, China*, pages 4976–4980, 2020.

[Quintas 2022c] Quintas S., Mauclair J., Woisard V. et Pinquier J., Automatic Assessment of Speech Intelligibility using Consonant Similarity for Head and Neck Cancer, dans *23rd INTERSPEECH Conference : Human and Humanizing Speech Technology (INTERSPEECH 2022)*, pages 3608–3612, Incheon, South Korea, Septembre 2022c, URL https://hal.science/hal-03716420.

[Quintas 2024] Quintas S., Vaysse R., Balaguer M., Roger V., Mauclair J., Farinas J., Woisard V. et Pinquier J., Sami : an m-health application to telemonitor intelligibility and speech disorder severity in head and neck cancers, *Frontiers in Artificial Intelligence*, 7 :1359094, 2024.

[Ravi 2019] Ravi V., Park S. J., Afshan A. et Alwan A., Voice quality and between-frame entropy for sleepiness estimation, dans *Interspeech 2019*, pages 2408–2412, 09 2019.

[Robert 2011] Robert M., Analyse acoustique des troubles articulatoires chez les patients atteints de paralysie faciale périphérique, Rapport technique, Ecole d'Orthophonie de Paris VI, 2011.

[Rossi 1998] Rossi M. et Peter-Defare E., Les lapsus ou comment notre fourche a langué, *Presses Universitaires de France*, 1998.

[Rouas 2004] Rouas J.-L., Farinas J. et Pellegrino F., Evaluation automatique du débit de la parole sur des données multilingues spontanées, *XXVe Journees d'Etude sur la Parole (JEP 2004), Fes, Maroc*, pages 437–440, 2004.

[Roulstone 2015] Roulstone S., Exploring the relationship between client perspectives, clinical expertise and research evidence, *International journal of speech-language pathology*, 17 :1–11, 04 2015.

[Rumbach 2019] Rumbach A. F., Finch E. et Stevenson G., What are the usual assessment practices in adult non-progressive dysarthria rehabilitation ? a survey of australian dysarthria practice patterns, *Journal of Communication Disorders*, 79 :46–57, 2019, URL https://www.sciencedirect.com/science/article/pii/S0021992418300595.

[Sahraoui 2022] Sahraoui H., Baqué L., Mauclair J. et Martínez-Ferreiro S., A corpus-based study of pauses and dysfluencies in autobiographic discourse and picture description of individuals with non-fluent aphasia, dans *International Conference - Science of Aphasia Conference Meeting*, volume 27, pages 61–64, Stem-, Spraak- En Taalpathologie, University of Bordeaux, Septembre 2022.

[Sahraoui 2015] Sahraoui H., Mauclair J., Baqué L. et Nespoulous J.-L., What do pause patterns in non-fluent aphasia tell us about monitoring speech ? a study of morph-syntactic complexity, accuracy and fluency in agrammatic sentence and connected discourse production., dans *53rd Annual Meeting of Academy of Aphasia (2015)*, Tucson, United States, Octobre 2015.

[San-Segundo 2001] San-Segundo R., Pellom B., Hacioglu K., Ward W. et Pardo J., Confidence measures for spoken dialogue systems, dans *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Salt Lake City, USA, Mai 2001.

[Schiller 2006] Schiller N. O., Phonological encoding in speech production., dans *ExLing*, pages 53–60, 2006.

[Sinha 2011] Sinha I. P., Smyth R. L. et Williamson P. R., Using the delphi technique to determine which outcomes to measure in clinical trials : Recommendations for the future based on a systematic review of existing studies, *PLOS Medicine*, 8(1) :1–5, 01 2011.

[Sisman 2020] Sisman B., Yamagishi J., King S. et Li H., An overview of voice conversion and its challenges : From statistical modeling to deep learning, *IEEE/ACM Trans. Audio, Speech and Lang. Proc.*, 29 :132 ?157, nov 2020, URL https://doi.org/10.1109/TASLP.2020.3038524.

[Snyder 2018a] Snyder D., Garcia-Romero D., Sell G., Povey D. et Khudanpur S., X-vectors : Robust dnn embeddings for speaker recognition, dans *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5329–5333, 2018a.

[Snyder 2018b] Snyder D., Garcia-Romero D., Sell G., Povey D. et Khudanpur S., X-vectors : Robust dnn embedings for speaker recognition, *Proceedings of ICASSP*, 2018b.

[van Son 1996] van Son R. et Pols L., An acoustic profile of consonant reduction, dans *Psychosomatic Medicine*, Janvier 1996.

[Tachioka 2013] Tachioka Y., Hanazawa T. et Iwasaki T., Dereverberation method with reverberation time estimation using floored ratio of spectral subtraction, *Acoustical Science and Technology*, 34(3) :212–215, 2013.

[Tomimoto 2008] Tomimoto J. et Takaoka Y., Le français, une langue imprononçable pour les japonais ?, *Rencontres Pédagogiques du Kansas*, 2008.

[Tremblay 2017] Tremblay P., Sato M. et Deschamps I., Age differences in the motor control of speech : An fMRI study of healthy aging, *Human Brain Mapping*, 38(5) :2751–2771, Mai 2017, URL https://hal.archives-ouvertes.fr/hal-03371899.

[Turan 2020] Turan M. A. T., Vincent E. et Jouvet D., Achieving Multi-Accent ASR via Unsupervised Acoustic Model Adaptation, dans *INTERSPEECH 2020*, Shanghai, China, Octobre 2020, URL https://hal.inria.fr/hal-02907929.

[Turoff 1996] Turoff M. et Hiltz S. R., Computer-based delphi processes, dans Ziglio M. A. . E., rédacteur, *Gazing into the oracle : The Delphi method and its application to social policy and public health*, pages 56–85, Jessica Kingsley Publishers, 1996.

[Van Lierde 2012] Van Lierde K., Browaeys H., Corthals P., Mussche P., Van Kerkhoven E. et De Bruyn H., Comparison of speech intelligibility, articulation and oromyofunctional behaviour in subjects with single-tooth implants, fixed implant prosthetics or conventional removable prostheses, *JOURNAL OF ORAL REHABILITATION*, 39(4) :285–293, 2012, URL http://dx.doi.org/10.1111/j.1365-2842.2011.02282.x.

[van Son 1999] van Son R. et Pols L. C., An acoustic description of consonant reduction1this paper is an extended version of a paper presented at icslp '96 in philadelphia (van son and pols, 1996).1, *Speech Communication*, 28(2) :125–140, 1999, URL https://www.sciencedirect.com/science/article/pii/S0167639399000096.

[Vasilescu 2004] Vasilescu I., Candea M. et Adda-Decker M., Hésitations autonomes dans 8 langues : étude acoustique et perceptive, dans *Colloque MIDL, Modélisations pour l'identification des langues et des variétés dialectales*, Novembre 2004.

[Vaswani 2017a] Vaswani A., Parmar N. S. N., Uszkoreit J., Jones L., Gomez A. N., ?ukasz Kaiser et Polosukhin I., Attention is all you need, *31st Conference on Neural Information Processing System*, 2017a.

[Vaswani 2017b] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser L. et Polosukhin I., Attention is all you need, dans *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000 ?6010, Curran Associates Inc., Red Hook, NY, USA, 2017b.

[Vaysse 2023] Vaysse R., *Caractérisation automatique du rythme de la parole : application aux cancers des voies aéro-digestives supérieures et à la maladie de Parkinson*, Thèse de doctorat, EDMITT, 2023, URL http://www.theses.fr/2023TOU30062/document.

[Veselỳ 2013] Veselỳ K., Ghoshal A., Burget L. et Povey D., Sequence-discriminative training of deep neural networks., dans *Interspeech*, volume 2013, pages 2345–2349, 2013.

[Vipperla 2008] Vipperla R., Renals S. et Frankel J., Longitudinal study of ASR performance on ageing voices, dans *INTERSPEECH*, pages 2550–2553, 2008.

[von der Gracht 2012] von der Gracht H. A., Consensus measurement in delphi studies : Review and implications for future quality assurance, *Technological Forecasting and Social Change*, 79(8) :1525–1536, 2012.

[Walsh 2005] Walsh R., Meaning and purpose : A conceptual model for speech pathology terminology, *Advances in Speech Language Pathology*, 7(2) :65–76, 2005, URL https://doi.org/10.1080/14417040500125285.

[Walsh 2006] Walsh R., A history of the terminology of communication sciences and disorders., *Australian Federal Government Department of Education, Science and Training.*, 2006.

[Wang 2022] Wang D., Liu S., Wu X., Lu H., Sun L., Liu X. et Meng H., Speaker identity preservation in dysarthric speech reconstruction by adversarial speaker adaptation, dans *2022 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6677–6681, 05 2022.

[Wang 2018] Wang Z., Zhang J. et Xie Y., L2 mispronunciation verification based on acoustic phone embedding and siamese networks, dans *2018 11th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pages 444–448, Nov 2018.

[Wessel 2005] Wessel F. et Ney H., Unsupervised training of acoustic models for large vocabulary continous speech recognition, *IEEE Transactions on Speech and Audio Processing*, 13 :23–31, 2005.

[Witt 2000] Witt S. M. et Young S. J., Phone-level pronunciation scoring and assessment for interactive language learning, *Speech communication*, 30(2-3) :95–108, 2000.

[Woisard 2021] Woisard V., Astésano C., Balaguer M., Farinas J., Fredouille C., Gaillard P., Ghio A., Giusti L., Laaridh I., Lalain M., Lepage B., Mauclair J., Nocaudie O., Pinquier J., Pouchoulin G., Puech M., Robert D. et Roger V., C2SI corpus : a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers, *Language Resources and Evaluation*, 55(1) :173–190, 2021, URL https://doi.org/10.1007/s10579-020-09496-3.

[Wolff 2018] Wolff M., Vanderhaegen F., Brethault M., Brisson H. et Mollard R., Vers une possible compréhension de l ?effet tunnel : une étude exploratoire, dans *Ergo'IA 2018 conference*, pages 3–5, 10 2018.

[Yamasaki 1999] Yamasaki H. et Halle P., How do native speakers of japanese discriminate and categorize french /r/ and /l/ ?, *Proceedings of ICPhS*, pages 909–912, 1999.

[Young 1994] Young S., The HTK hidden markov model toolkit : Design and philosophy, *Entropic Cambridge Research Laboratory, Ltd*, 2 :2–44, 01 1994.

[Zargarbashi 2019] Zargarbashi S. et Babaali B., A multi-modal feature embedding approach to diagnose alzheimer disease from spoken language, *arXiv :1910.00330*, 2019.