Modélisation de l'indice de sévérité du trouble de la parole à l'aide de méthodes d'apprentissage profond : d'une modélisation à partir de quelques exemples à un apprentissage auto-supervisé via une mesure entropique

Vincent Roger

▶ **To cite this version:**

Vincent Roger. Modélisation de l'indice de sévérité du trouble de la parole à l'aide de méthodes d'apprentissage profond : d'une modélisation à partir de quelques exemples à un apprentissage auto-supervisé via une mesure entropique. Apprentissage [cs.LG]. Université Paul Sabatier - Toulouse III, 2022. Français. NNT : 2022TOU30180 . tel-03935738

HAL Id: tel-03935738
https://theses.hal.science/tel-03935738

Submitted on 12 Jan 2023

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

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

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par : *l'Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)*

Présentée et soutenue le *29/09/2022* **par :**

Vincent ROGER

Modélisation de l'indice de sévérité du trouble de la parole à l'aide de méthodes d'apprentissage profond

D'une modélisation à partir de quelques exemples à un apprentissage auto-supervisé via une mesure entropique

---

JURY

Hervé Glotin	Professeur d'Université	Président du Jury
Cécile FOUGERON	Professeure d'Université	Rapporteure
Jean-François BONASTRE	Professeur d'Université	Rapporteur
Julien PINQUIER	Maître de Conférences	Directeur de thèse
Jérome FARINAS	Maître de Conférences	Co-encadrant de thèse
Virginie WOISARD	Professeure d'Université	Co-directrice de thèse

---

École doctorale et spécialité :
MITT : Image, Information, Hypermédia

Unité de Recherche :
Institut de Recherche en Informatique de Toulouse (UMR 5505)

Directeur(s) de Thèse :
Julien PINQUIER, *Jérome FARINAS* et *Virginie WOISARD*

Rapporteurs :
Cécile FOUGERON et *Jean-François BONASTRE*

Remerciements

Je tiens à remercier mes directeurs de thèses, Jérôme Farinas, Julien Pinquier et Virginie Woisard. Ils ont été présents tout le long de ce parcours. Je tiens particulièrement à exprimer ma gratitude à Jérôme et Julien pour m'avoir soutenue pendant les moments difficiles. En particulier durant la pandémie où mon moral était le plus bas. Je remercie également Gauthier Arcin pour son travail exemplaire durant son CDD auprès de moi.

Je suis reconnaissant envers mes deux rapporteurs Cécile Fougeron et Jean-François Bonastre pour leurs relectures, ainsi que de m'avoir permis de réaliser la version que vous allez lire de ce manuscrit. Je suis également reconnaissant envers Hervé Glotin pour avoir présidé ma soutenance de doctorat.

Je remercie ma femme Christelle Duchossoy tout comme mes amis pour leur soutien. Je remercie tous les membres de l'équipe SAMOVA pour leur accueil chaleureux, c'est une équipe remplie de personnes bienveillantes et de bons conseils. Les moments passés les midis et en soirée avec les doctorants et ingénieurs de recherches ont été merveilleux. Parmi ces doctorants, j'ai trouvé bien plus que des collègues, d'excellents amis ! Je remercie également les membres du projet RUGBI avec qui j'ai pu échanger de mes avancées et profiter d'un environnement pluridisciplinaire pour ma thèse.

Enfin, je tiens à finir par être reconnaissant envers ma famille et mes parents pour les valeurs qu'ils m'ont transmises.

Je dédie cette thèse à Jérôme Farinas et Julien Pinquier
sans qui elle n'aurait jamais vu le jour.
Mais également à mon amour,
Christelle Duchossoy qui m'a encouragée tout le long.

Table des matières

Table des figures

Introduction générale

Contexte et enjeux médicaux

De nombreuses pathologies peuvent altérer la production de la parole [Enderby, 2013] : elles peuvent être liées à des atteintes neurologiques comme la maladie de Parkinson, ou bien aux conséquences des cancers et de leurs traitements (opérations, radiothérapie et/ou chimiothérapies). Entre 2007 et 2016, les cancers de la cavité buccale et du pharynx représentaient 3% des cancers diagnostiqués chaque année aux États-Unis [Ellington, 2020], et leur incidence était en augmentation sur cette période. Il s'agit de pathologies qui, par leur localisation, ont une grande influence sur la qualité de vie des patients [Walshe and Miller, 2011], car les fonctions de communication sont directement affectées, et cela génère une grande gêne dans la vie au quotidien. Le suivi des patients repose sur une évaluation perceptuelle de la qualité de production de leur parole. L'introduction de mesures automatiques présenterait des avantages pour les soignants, puisque cela apporterait plus de rapidité dans la production d'un indice caractérisant l'atteinte de la production de parole. En effet, il est nécessaire d'amalgamer les annotations perceptuelles d'un groupe d'experts afin de produire des annotations fiables [Woisard and Lepage, 2010]. Cela permet de pallier au phénomène d'habituation du praticien auprès de la voix de son patient, mais également de réduire la variabilité issue de ces mesures.

Les praticiens produisent classiquement plusieurs types de mesures : une mesure de la sévérité du trouble de la parole, mais également une mesure de l'intelligibilité. Ces termes peuvent avoir des interprétations différentes au sein de la communauté scientifique. Cela a fait l'objet d'une étude pour essayer de confronter les définitions et les différents cas d'usage et d'en obtenir un consensus [Pommée et al., 2021]. Dans la suite de ce document, nous utilisons la définition suivante pour l'intelligibilité (issue de ces travaux) : « L'intelligibilité fait référence à la reconstruction d'un énoncé au niveau acoustico-phonétique, l'information liée à l'intelligibilité est donc portée par le signal acoustique (c.-à-d. l'intelligibilité se concentre sur l'information dépendante du signal). Cette reconstruction est rendue possible à la fois par les capacités de production phonétique-acoustique du locuteur et par les capacités de décodage acoustico-phonétique de l'auditeur. ». Tandis que pour la sévérité du trouble de la parole, nous la définissons comme une mesure de l'impact de la maladie sur la production de parole du patient [Woisard and Lepage, 2010]. Cette mesure essaie de ne pas transcrire les capacités de décodage acoustico-phonétique de l'auditeur, mais l'impact de la maladie sur les capacités de production phonétique-acoustique du locuteur. Les deux mesures sont par définition corrélées : plus la parole d'un patient est sévèrement atteinte, plus son score d'intelligibilité de production

de la parole est affectée [Kent, 1992]. Enfin, la sévérité mesure des altérations non décelables par l'intelligibilité telles que des paramètres temporels et/ou prosodiques [Auzou et al., 2007]. Par souci de lisibilité, nous simplifierons ces indices en parlant d'indice d'intelligibilité et d'indice de sévérité de la parole. Il est à noter qu'il existe dans la communauté « parole » d'autres indices dits d'intelligibilité. Cependant, ils mesurent l'impact de l'environnement sur l'information véhiculée par un orateur et non pas la capacité de l'orateur à communiquer.

L'inconvénient d'indices comme l'intelligibilité de la parole ou la sévérité de la parole est qu'ils sont fondés sur une évaluation perceptuelle. De ce fait, les mesures intra-experts et inter-experts ne sont pas suffisamment fiables et reproductibles. De plus, une instabilité d'évaluation peut provenir des phénomènes d'accoutumance dus à l'expérience du thérapeute des troubles de la parole en général, mais également des altérations de la parole de leurs patients. Ainsi, il est impossible d'avoir une mesure stable avec un seul expert. Pour réaliser une mesure stable et non biaisée de ces deux indices, il est nécessaire de réunir un groupe d'experts [Woisard and Lepage, 2010]. Ceci est irréalisable en pratique. Les experts ne peuvent pas se réunir régulièrement pour chaque patient. Par conséquent, il n'est pas envisageable de déplacer un groupe d'experts pour évaluer l'évolution de ces indices au domicile des patients.

Ainsi, un système de mesure automatique établi sur l'écoute d'un patient permettrait de mieux refléter l'impact de la maladie sur la qualité de vie des personnes atteintes. Les patients pourraient, par exemple, réaliser ces mesures à l'aide d'applications mobiles et cela fournirait plus d'information aux experts voulant s'occuper de ces derniers. De plus, avoir une mesure proche d'un groupe d'experts permettrait le suivi plus régulier d'un patient. Ce dernier point libérerait du temps aux experts pour se focaliser sur un meilleur accompagnement de leurs patients.

Problématique et solutions travaillées

En étudiant les approches précédentes à ma thèse (dont la revue est présente dans le chapitre 2 de la première partie de ce manuscrit), nous avons constaté qu'il n'y avait pas de méthode utilisant les avantages des réseaux profonds : leur fort pouvoir de représentation, leur robustesse et leurs résultats souvent au niveau de l'état de l'art en parole. Par robuste, nous entendons une robustesse aux bruits environnementaux, aux accents des locuteurs et à l'évolution de la condition du locuteur (plus de détails sur ces besoins sont donnés dans le chapitre 2 de la seconde partie). Les résultats de ces approches en parole pathologique précèdent mes travaux ne sont pas adéquats pour envisager leur utilisation en milieux médicaux (nous détaillerons ceci dans le chapitre 2).

De cet état des lieux, émerge donc la problématique suivante : est-ce que les techniques d'apprentissage peuvent, avec un corpus de données limité, modéliser le concept d'indice de sévérité du trouble de la parole tout en étant robustes ? Et, si oui, serait-il possible d'envisager une application médicale à ces travaux ?

C'est dans l'environnement décrit en chapitre 1 de la première partie que j'ai pu évoluer pour répondre à cette question. Possédant un corpus de taille restreinte (environ une heure

d'enregistrements de lecture au total), comment l'utiliser convenablement pour réaliser notre modélisation du concept de sévérité ? Ainsi, nous commencerons par voir en première partie le cadre scientifique dans lequel mon travail a commencé pour constater la quantité, la qualité et le type de données que nous possédons dans le chapitre 3. L'émergence des réseaux de neurones (utilisant de l'apprentissage profond, généralement gourmand en ressources) a permis des avancées fulgurantes dans plusieurs tâches de paroles telles que la reconnaissance automatique de la parole, la reconnaissance du locuteur ou la reconnaissance d'émotions. Ainsi, est-ce que de telles technologies peuvent modéliser notre problème ? Pour répondre à cette question, nous montrerons nos travaux d'apprentissage supervisé dans le chapitre 4. Dans celui-ci, nous avons exploré l'apprentissage de la mesure de sévérité (plus précisément, le score moyen donné par les experts pour chaque participation) par un modèle. Dans ce travail, nous avons essayé de définir quelles données sont nécessaires et en quelles quantités pour entraîner et évaluer notre système. Ces modèles ont été appris à l'aide d'approches supervisées (apprentissage guidé par la vérité terrain) et utilisant des techniques adaptées aux corpus avec une quantité de données limitée.

Ayant un corpus restreint en quantité de données, ces approches prometteuses et ayant des résultats intéressants nous semblent limitées pour apprendre une mesure robuste et corrélée au score de sévérité. De ce fait, en chapitre 5 de la seconde partie, nous nous sommes tournés vers des approches autosupervisées, c'est-à-dire dont l'apprentissage est guidé par les données et leurs caractéristiques plutôt que par la vérité terrain des experts. Ainsi, ce type de modèle nous a permis de créer une mesure fortement corrélée à l'indice de sévérité (avec une corrélation de Spearman de -0,86). De plus, le type de modèle utilisé assure une certaine robustesse de l'approche et nous a permis d'utiliser la totalité du corpus cancer pour évaluer notre système. Nos résultats nous ont encouragés à envisager la création d'une application mobile que nous décrirons dans le chapitre 6. Cette dernière est maintenant déployée au sein du CHU de Toulouse.

Avant d'en voir plus sur cette application, commençons par le cadre scientifique de mon travail.

Première partie

Cadre scientifique

Introduction de partie

Ce manuscrit de thèse a été découpé en deux parties : cette première partie décrit le cadre scientifique dans lequel j'ai pu évoluer pour mon travail et la seconde se concentre plus particulièrement sur les contributions. Le financement de mon doctorat est issu d'un financement conjoint entre l'Université Fédérale de Toulouse et la région Occitanie, et s'intercale dans un projet de recherche ANR RUGBI. Le chapitre 1 détaille l'environnement scientifique dans lequel mon travail de thèse s'inscrit. Ainsi, les projets réalisés en amont de ma thèse et le contexte de travail auquel j'ai eu accès sont décrits.

À la suite de cela, le chapitre 2 présente un état de l'art des mesures automatiques essayant de reproduire les scores de sévérités et d'intelligibilité de production précédant mon travail de thèse. Cet état des lieux se limite aux mesures semblables à l'intelligibilité et sévérité de production et ne prendra pas en compte celles pour lesquelles les bruits environnementaux impactent la mesure.

Enfin, dans le chapitre 3, nous décrivons le corpus auquel j'ai eu accès. J'ai effectué un travail d'analyse quantitative des données afin de vérifier les équilibres suivant diverses dimensions du corpus cancer. J'ai ensuite produit un travail de préparation des données, de normalisation et de correction d'erreurs. Il s'agit, en effet, d'une étape indispensable en vue d'une contribution portant sur de la modélisation automatique.

1

Environnement multidisciplinaire

Ce projet de thèse fait suite aux travaux réalisés par les chercheurs membres du projet INCA C2SI [1]. Leur objectif était de démontrer que l'indice de sévérité de la parole carcinologique (C2SI) obtenu par un outil de traitement automatique de la parole donne des résultats équivalents ou supérieurs à un score d'intelligibilité de la parole obtenu par des auditeurs humains. Pour ce faire, les chercheurs de ce projet ont réalisé le recueil de données de parole de participants que nous retrouvons dans notre corpus cancers, dont la description détaillée se trouve en chapitre 3 et leur meilleur score automatique est détaillé en fin de chapitre 2.

Suite à ce projet, mon doctorat a été financé par la région Occitanie ainsi que par l'université fédérale de Toulouse au travers de la bourse N° 2018-1290 (ALDOCT No500).

L'objectif de ce doctorat est de réaliser un système de mesure automatique de la sévérité de la parole. Cela part du constat qu'il manque des méthodes uniformisées pour évaluer des résultats de gêne de production de la parole pathologique. Cet indice de handicap de parole est fortement corrélé (entre 0,7 et 0,9) avec le domaine de la parole du questionnaire de qualité de vie utilisé par les praticiens [Borggreven et al., 2007, Dwivedi et al., 2011, Thomas et al., 2009]. Ainsi, en plus de refléter la qualité de vie des patients, la création d'une mesure automatique de la parole permettrait :

- de compléter l'expression des résultats thérapeutiques par des indices de pronostic fonctionnel,
- de décrire le profil des patients en fonction du niveau de risque du handicap de parole,
- d'ajuster les gestes thérapeutiques pour réduire leurs conséquences fonctionnelles,
- d'étudier les rapports entre le déficit anatomo-fonctionnel et les indices de déficit de parole et de communication.

Les hôpitaux ne disposant pas nécessairement de chambre anéchoïque, la solution apportée doit proposer une certaine robustesse à l'environnement de l'hôpital. De plus, ces derniers nécessitent d'avoir un accès à un outil de traitement automatique de la parole permettant un haut niveau de fiabilité et d'objectivité.

Pour résoudre ce problème, j'ai souhaité modéliser l'indice de sévérité avec des approches utilisant des réseaux profonds **et** adaptés à peu de données. J'ai également souhaité modéliser l'indice de sévérité grâce à des modèles auto supervisés (dont la supervision vient des données

1. Plus de détails sur ce projet sont disponibles ici : https://www.irit.fr/SAMOVA/site/projects/previous/c2si/

et non d'une vérité terrain réalisée par des experts). Enfin, j'ai souhaité réaliser une application pouvant être utilisée en milieu médical.

En parallèle de ce travail, dans mon équipe de recherche (dont certains membres font partie du groupe de recherche européen TAPAS [2]), des doctorants sont partis sur les sujets suivants :

- l'évaluation de la pertinence clinique des mesures d'intelligibilité de la parole dans un contexte oncologique.
- la mesure d'altération de la communication, l'analyse automatique la parole spontanée et la prédiction de l'altération de la communication par les paramètres acoustiques extraits automatiquement.
- la modélisation automatique du rythme de la parole par la recherche d'une métrique rendant compte des différences de fluence prosodique entre des personnes saines et des personnes atteintes de pathologies affectant leur production orale.
- la modélisation de l'indice d'intelligibilité à l'aide d'approches basées sur des représentations (par exemple les x-vectors dont nous comparerons les résultats avec les nôtres dans le chapitre 5).

J'ai également participé au projet de recherche RUGBI (financé par l'agence nationale de recherche française [3]). Ce projet regroupe des membres de l'Institut de Recherche en Informatique de Toulouse (IRIT), du CHU de Toulouse, de l'Université Toulouse Jean Jaurès (UT2J) du Laboratoire Informatique d'Avignon (LIA) et du Laboratoire Parole et Langage d'Aix en Provence (LPL). La liste des participants se retrouve sur la page du projet [4]. L'objectif du projet RUGBI est de compléter les outils des thérapeutes pour les aider à réaliser une prise en charge non invasive, rapide et abordable. Pour ce faire, le projet a été subdivisé en trois lots de travaux auxquels j'ai pu participer lors de mon doctorat :

1. Gestion et sélection des données,
2. Identification des unités et des tâches pertinentes pour évaluer l'intelligibilité dans les troubles de production de la parole,
3. Production de modélisations et mesures automatiques.

J'ai principalement normalisé les données : nommage, correction des erreurs de vérité terrain et fouille des archives. Il en résulte la définition d'une base de données sur les patients atteints de cancer que je présenterai en détail dans le chapitre 3. Pour le second lot, mes contributions liées aux approches supervisées indiquent quelles unités phonémiques permettent une modélisation plus aisée du concept de mesure perceptive de sévérité : c'est-à-dire quelles unités peuvent être pertinentes pour la construction de futurs textes visant à évaluer l'indice de sévérité du participant. Enfin, mon doctorat étant porté sur la modélisation du concept de sévérité, j'ai bien entendu participé activement au dernier lot. Ainsi, j'ai bénéficié de cet environnement multidisciplinaire pour tirer des leçons de chacun et pour partager mes avancées.

Maintenant que mon environnement de thèse est exposé, nous allons dans le chapitre suivant, dégager un état de l'art des différentes techniques permettant une mesure automatique de l'intelligibilité et de la sévérité de la production de la parole.

2. https://www.tapas-etn-eu.org/
3. ANR-18-CE45-0008
4. https://www.irit.fr/rugbi/consortium/

2

Mesures automatiques d'intelligibilité et de sévérité

La majorité des mesures automatiques de production de la parole, que nous avons recensée en début de thèse, sont fondées sur la reproduction de l'intelligibilité de production. Ainsi, lorsque nous regardons les mesures automatiques d'intelligibilité de production de la parole (comparable à la définition de [Pommée et al., 2021] vue en introduction générale), nous remarquons que l'ensemble des méthodes utilisent une approche supervisée pour modéliser le concept d'intelligibilité. C'est-à-dire que la vérité terrain est utilisée pour construire un modèle permettant de prédire un score ou une catégorie à partir de l'enregistrement audio d'un patient.

Les méthodes examinées utilisaient soit des corpus avec une vérité terrain représentant des catégories simples (la personne est intelligible ou non), soit un score borné (souvent la moyenne de scores entiers, donnés par un jury d'experts). Certaines méthodes disposant d'un score borné ont modifié le problème en créant des catégories (par exemple : non intelligible, peu intelligible, intelligible). Ainsi, nous pouvons définir les approches recensées en deux catégories :

- les approches reproduisant un score continu entre deux bornes, c'est-à-dire les approches de régression,
- les approches catégorisant un participant, où chaque catégorie représente un intervalle de score continu.

La première est une mesure au plus proche de ce que peut produire un jury d'expert, tandis que la seconde est plus proche de ce qu'un écouteur « naïf » peut produire. La catégorisation de production d'un locuteur ne permet pas de différencier les productions appartenant à une même catégorie. Ceci limite la finesse du suivi de la progression d'un patient suivant un traitement et/ou une rééducation (modulo la finesse des catégories utilisées). Néanmoins, la catégorisation des productions de locuteur faciliterait le choix de protocole de rééducation appliqué au patient (en supposant qu'une approche différente soit utile en fonction de la catégorie associée au locuteur). Il est cependant difficile de définir les catégories supérieures au cas binaire patient/non-patient. En effet, il n'existe pas (à notre connaissance) de consensus sur les catégories de locuteurs. De plus, si nous définissons des catégories à partir d'intervalles de scores, il est ardu de différencier des locuteurs proches de la frontière de deux catégories.

Pour ces deux approches, les entrées des modèles peuvent être identiques, leurs sorties sont différentes. Par conséquent, l'évaluation de ces approches est différente. Pour les approches de régression, une corrélation de Pearson ([Benesty et al., 2009], notée ρ) entre les scores prédits par le modèle et les scores produits par les experts permet d'évaluer les résultats de tels modèles. Tandis que pour les approches de classification, l'utilisation d'une métrique de précision (avec cette métrique égale au nombre d'éléments bien prédits sur le nombre total d'exemples) entre le label prédit et le label des experts définit le pourcentage de bonnes précisions (avec le taux d'erreur étant égal à $1 -$ précision).

Les méthodes recensées ne fournissent généralement pas leur base de données et rarement leur code d'évaluation : il est alors impossible de les comparer finement [5]. Bien que nous puissions reproduire la plupart des approches recensées [6], il nous semble plus intéressant de définir des critères objectifs permettant une évaluation/comparaison des approches. Ainsi, pour chaque approche, nous évaluons :

- la complexité de la tâche. Cela peut se traduire par le contenu de la base de données (par exemple, les conditions d'enregistrement) ou le type de mesure produite. En effet, nous supposons qu'une classification binaire (en patient/non patient) est plus simple que de reproduire une mesure d'intelligibilité/sévérité évaluée sur une échelle de 0 à 100.
- la robustesse aux bruits. Cela se traduit par soit une base de données contenant des bruits environnementaux, soit par l'utilisation de mécanismes de débruitage.
- la robustesse au locuteur. Le modèle doit être capable, pour un même locuteur, de donner des scores similaires lorsque les scores de sévérité (ou d'intelligibilité) sont similaires entre deux enregistrements. De plus, lorsqu'un même locuteur produit des enregistrements avec des scores différents, le modèle doit également être capable de prédire des scores semblables aux scores d'experts. Pour vérifier une telle robustesse, il est nécessaire d'avoir plusieurs enregistrements pour l'ensemble des locuteurs utilisés pour évaluer le système.
- la robustesse aux accents locaux. Nous considérons qu'une approche adaptée à une langue d'un pays ne doit pas être perturbée par une personne possédant un accent. Cela peut se traduire par l'utilisation d'un modèle global (appris sur une large variété d'accents) ou de caractéristiques du signal testé sur une base contenant une large variété d'accents.

Nous avons considéré pour ce travail de thèse que si une approche possède les trois derniers points, cette méthode peut être envisagée pour une utilisation en milieu médical.

2.1 Classification de production de parole

En classification par intervalles de scores, nous avons recensé le travail de [Fang et al., 2017]. Ces derniers ont testé leur approche sur deux corpus de lecture de mots isolés allemands : le premier corpus comporte uniquement des patients (55) [Clapham et al., 2012]

5. Par évaluation fine, nous entendons une évaluation par une même métrique d'évaluation sur une même base de données avec accès aux données pour évaluer les forces et faiblesses de chaque approche.

6. Avec quelques limites. En effet, nous possédons des bases de données différentes et nous ne connaissons pas certains hyperparamètres utilisés pour apprendre les modèles.

alors que le second comprend des participants contrôles (650, en complément des 1320 patients) [Martínez et al., 2012]. Il est à noter que ces corpus ne sont pas utilisés dans le milieu médical et sont composés d'une phrase seulement. De plus, ces phrases ne représentent pas l'étendue des phonèmes allemands. Le fait de ne pas représenter tous les phonèmes pouvant être produits ne nous paraît pas idéal pour évaluer l'intelligibilité, car chaque phonème peut contribuer au concept de manière différente. Si cela est le cas, nous imaginons qu'il y a des spécificités en fonction de la maladie des patients et qu'une citation des auteurs aurait été faite dans ce sens. Dans les deux cas, il s'agissait d'une classification binaire : « voix saine » *vs* « voix non saine ».

Leur approche consiste à extraire des paramètres du signal puis à modéliser les deux classes. Les auteurs ont proposé des paramètres proches des coefficients cepstraux (MFCC) ainsi qu'un large ensemble de paramètres liés à l'acoustique et à la théorie du chaos. Il est à noter que les auteurs ont comparé leurs paramètres aux MFCC et obtiennent des résultats similaires bien que légèrement inférieurs. Les paramètres acoustiques portaient sur la fréquence fondamentale, sur la qualité sonore (comme le jitter et le shimmer) et sur des caractéristiques spectrales (comme les centroïdes spectraux, l'entropie spectrale, etc.). Les paramètres liés à la théorie du chaos sont également utilisés par les auteurs pour analyser le conduit vocal des patients. Ainsi les auteurs de cette approche ont choisi les paramètres suivants : le plus grand exposant de Lyapunov, l'entropie approximative et la complexité de Lempel-Ziv [Abásolo et al., 2015]. Ils ont ensuite proposé une méthode de projection de ces paramètres pour la présenter à une Machine à Vecteurs de Support (SVM). Ils obtiennent une précision de 75,87% pour le corpus NKI-CCRT [Clapham et al., 2012] et de 78,97% pour le corpus [Martínez et al., 2012]. Les résultats obtenus ne sont pas satisfaisants pour une utilisation médicale. En effet, aucune robustesse aux locuteurs, aux bruits environnementaux ou aux accents locaux n'a été testée (mise en place). De plus, il faut mitiger leurs résultats, puisque pour une tâche binaire leurs ensembles d'évaluations sont déséquilibrés. Ainsi, prédire systématiquement qu'un patient est non intelligible sur le corpus NKI-CCRT donne une précision de 54,29% (avec 341 échantillons intelligibles et 405 non intelligibles dans l'ensemble d'évaluation) et prédire qu'une personne est un patient donne une précision de 51,59% (avec des échantillons provenant de 198 patients et de 211 contrôles). Ainsi, leurs résultats restent supérieurs à une prédiction mono-label. Nous ne retenons pas leur approche pour la suite de nos travaux, car ils n'utilisent pas des caractéristiques du signal qui ont été reprises dans la littérature et les auteurs n'ont pas mesuré l'impact individuel de chacune des caractéristiques qu'ils proposent (comparé à l'utilisation des MFCC seuls).

2.2 Régression de scores d'experts

Pour les approches de régression du score de perception, nous avons recensé bien plus d'approches. Ces dernières sont résumées en table 2.1.

Ainsi, nous retrouvons des corpus de lecture de texte français [Fougeron et al., 2010] (intelligibilité perçue échelonnée de 0 à 3) et allemand [Maier et al., 2009] (intelligibilité perçue échelonnée de 1 à 5). Nous retrouvons également de la lecture de mots isolés en anglais [Kim et al., 2008] (intelligibilité échelonnée de 0 à 100) et [Fletcher et al., 2017] (dont l'échelle du score n'est pas précisée). Seul un des papiers recensés ne précise ni la langue ni les mots utilisés

TABLE 2.1 – Table de compilation des approches de régression.

Corpus	Méthode	#Patients	#Contrôles	Résultat
[Kim et al., 2008]	[Janbakhshi et al., 2019]	15	13	$\rho = 0,94$
[Middag et al., 2008]	[Middag et al., 2008]	160	51	$\rho = 0,90$
[Maier et al., 2009]	[Maier et al., 2009]	100	40	NA [a]
[Fougeron et al., 2010]	[Laaridh et al., 2017]	99	30	$\rho = 0,88$
[Fletcher et al., 2017]	[Fletcher et al., 2017]	43		$\rho = 0,44$

a. Les auteurs fournissent une corrélation uniquement pour les patients de cancers oraux ($\rho = -0,9$) et pour les patients laryngectomisés ($\rho = -0,83$), sans donner de score global ni de score pour les contrôles.

dans leur base [Middag et al., 2008]. Sur l'ensemble de ces corpus recensés, aucun ne garantit l'utilisation de tous les phonèmes de la langue cible. Ceci nous semble être un biais de mesure, car il n'existe pas (à notre connaissance) d'étude sur ces langues démontrant que seulement certains phonèmes sont affectés par la maladie (en supposant que chaque maladie a le même impact, ce qui n'est même pas le cas pour des patients atteints d'une même maladie dont le stade et les traitements sont différents [Jacobi et al., 2010]). Il n'existe également pas (à notre connaissance) d'étude montrant que l'impact de la maladie est homogène pour l'ensemble des phonèmes (ce que nous supposons faux pour ce travail de thèse). Pour les corpus des méthodes recensées, les échelles d'évaluation de l'intelligibilité sont soit larges (échelle de 0 à 100), soit restreintes (2 à 5 valeurs possibles). Il est difficile d'évaluer la pertinence d'évaluation perceptive utilisant des échelles de valeurs larges, car le consensus d'experts sur des corpus comparables ont tendance à ne pas utiliser d'échelles de valeurs au-dessus de la dizaine de valeurs [Woisard and Lepage, 2010, Yamasaki et al., 2017, Martins et al., 2015]. Comparé au corpus cancer [Woisard et al., 2021], utilisé dans le travail lié à cette thèse, les valeurs des scores d'intelligibilité et de sévérité de production sont distribuées sur 11 valeurs (de 0 à 10) par les experts. Ce corpus utilise le score moyen des experts comme score de référence, ce qui nous donne un score dans l'intervalle $[0 - 10]$. De plus, certaines tâches de notre corpus (notamment la tâche de lecture) garantissent que les participants prononcent une assez bonne couverture de l'inventaire phonétique Français. Ceci nous permettra d'éviter les différentes erreurs de mesures recensées sur les corpus des méthodes revues dans ce chapitre et de faciliter la comparaison entre plusieurs méthodes.

Pour les méthodes que nous avons recensées, certaines utilisent la vraisemblance moyenne des mots ou des phonèmes prédits par un système de reconnaissance automatique de la parole (ASR). Un modèle de régression est alors construit sur ces vraisemblances [Middag et al., 2008, Maier et al., 2009]. Cette approche nécessite de connaître le texte ou les mots prononcés et de réaliser un alignement forcé sur les données. Néanmoins, ces approches n'ont pas mis en place des techniques permettant d'avoir une quelconque robustesse aux bruits environnementaux. De plus il n'y a pas plus de détails sur la robustesse au locuteur et aux accents locaux. Ceci étant nous supposons que leur approche possède une telle robustesse (bien qu'on ne peut la quantifier) vu que ces méthodes utilisent des modèles venant des systèmes de reconnaissance automatique de la parole. Pour la suite, nous nous sommes limités à ces deux références. En effet, leurs résultats sont les plus prometteurs des papiers recensés avec ce type d'approche et les résultats secondaires de Maier et al. sont intéressants pour la suite de nos travaux. Ainsi, ils

ont montré que, plus leur modèle de reconnaissance de la parole (utilisant des n-gram [7] comme modèle de langage) considérait le contexte (avec un n-gram plus élevé), plus leur approche se décorrélait de l'intelligibilité perçue. Ce résultat concorde avec la définition des mesures d'intelligibilité et de sévérité. Effectivement, si les évaluateurs connaissent le contexte, ces derniers ont tendance à surévaluer les scores des patients [Beukelman and Yorkston, 1980] et évaluent la compréhensibilité d'un patient. Les mesures comme l'intelligibilité et la sévérité sont censées être une composante de la compréhensibilité tout en excluant les connaissances globales que l'auditeur peut avoir sur la construction d'une phrase, l'accent, la thématique du message… Ainsi, ce phénomène se retrouve dans l'expérience de Maier et al. C'est par conséquent une indication sur les limitations du modèle que l'on doit garantir pour de telles approches.

Nous retrouvons également une méthode fondée sur les standards STOI [Falk et al., 2012] et ESTOI [Jensen and Taal, 2016]. Ces standards permettent de mesurer l'intelligibilité environnementale et nécessitent d'utiliser des enregistrements de référence non bruités pour les comparer à des enregistrements bruités. Ils supposent également que ces signaux soient alignés temporellement. Le STOI extrait les enveloppes temporelles des signaux vocaux non bruités et bruités en sous-bandes de fréquence. Les enveloppes sont ensuite soumises à une procédure d'écrêtage, comparées à l'aide de coefficients de corrélation linéaire à court terme et une prédiction finale de l'intelligibilité est construite comme une moyenne des coefficients de corrélation. Le ESTOI est une mesure fondée sur le STOI excepté qu'il calcule les coefficients de corrélation spectrale, qui sont finalement moyennés à travers le temps dans des segments d'analyse de 384 ms. Sur le même principe, [Janbakhshi et al., 2019] proposent d'utiliser des enregistrements contenant de la « parole saine » comme référence (en supposant que les conditions sonores restent les mêmes pour tous les enregistrements) en ayant des références pour chaque mot et d'utiliser les standards STOI et ESTOI pour calculer un score. Ainsi, par construction, il n'y a aucune robustesse aux bruits environnementaux. Nous émettons également des doutes quant à la robustesse au locuteur/accents d'une telle approche, car il n'y a pas d'indications des auteurs sur la sélection et utilisation des exemples de références (leur nombre, la distinction homme/femme, etc.). Néanmoins, leur utilisation de la parole « saine » comme référence nous à inspiré et c'est l'idée de base de notre approche (voir le sous-chapitre 5.2).

Nous avons également recensé une méthode de régression d'évolution de l'intelligibilité [Fletcher et al., 2017]. Les auteurs ont utilisé un outil automatique de régression (non précisé par les auteurs) sur les caractéristiques spectrales suivantes : le spectre moyen à long terme (LTAS), les spectres de modulation de l'enveloppe (EMS) et les MFCC. Les LTAS fournissent une représentation de l'information spectrale moyenne contenue dans le signal vocal de phrases entières. Cette caractéristique spectrale est opposée à l'EMS et aux MFCC qui sont calculés sur des fenêtres plus courtes. Ici, il est impossible d'évaluer l'approche des auteurs vu que nous ne disposons d'aucun élément sur le type de modèle utilisé. Ainsi, nous laissons de côté leur approche. Nous notons qu'une fois encore, les MFCC sont utilisés pour une mesure d'intelligibilité.

Enfin, la dernière méthode que nous avons recensée consiste à utiliser des i-vecteurs combinés à une régression par vecteur support (SVR) [Laaridh et al., 2017]. Nous pensons que cette

7. Les n-grams sont une séquence contiguë de n éléments d'un échantillon donné de parole et permettent de prédire le mot suivant d'une séquence de mots.

approche est la plus robuste envers les locuteurs, en raison de la nature des i-vecteurs (représentation de caractéristiques permettant de réaliser l'identification de locuteur). Néanmoins, nous n'avons aucune indication sur la robustesse d'environnements et de robustesse aux accents locaux (que nous supposons mauvaise, car l'accent peut faire partie des caractéristiques représentées par les i-vecteurs).

Bien que les résultats semblent globalement encourageants, nous devons rester prudents. D'une part, certains auteurs utilisent un corpus assez faible (15 patients [Janbakhshi et al., 2019]), ce qui laisse douter de la généralisation d'une telle méthode. D'autre part, l'approche de [Fletcher et al., 2017] n'est pas satisfaisante pour envisager une utilisation médicale. Néanmoins, plusieurs méthodes restent prometteuses [Middag et al., 2008, Maier et al., 2009, Laaridh et al., 2017].

Précédemment à mon travail de thèse, des essais avec les i-vecteurs et fondés sur des ASR ont été effectués sur notre corpus cancer [Balaguer et al., 2019]. Il en résulte que la meilleure approche a obtenu une corrélation de Spearman de 0,817 avec la sévérité pour la tâche de lecture sur tous les participants. Ainsi, ce travail nous sert de référence pour les expérimentations à suivre. Il est à noter que les auteurs de [Middag et al., 2008, Maier et al., 2009] précisent que ce genre d'approche empêche une prononciation différente pour les patients d'une autre région en raison des phonèmes fixes attendus (même si certains ajustements sont possibles et que cela est surtout le cas pour les modèles hybrides). De plus, les auteurs précisent que leur méthode n'est pas robuste au locuteur, c'est-à-dire qu'un même locuteur peut avoir des scores différents bien qu'ayant une prononciation équivalente.

Mes travaux de thèse prennent suite à ces mesures automatiques, mais avant de parler de mes contributions scientifiques, nous allons nous attarder sur l'analyse du corpus mis à ma disposition.

3

Présentation et analyse du corpus cancer

Le corpus de parole de personnes atteintes de cancer des voies aérodigestives supérieures (que nous simplifierons par « corpus cancer » pour la suite de la thèse) est un corpus médical composé d'enregistrements audio de 128 participants (patients et contrôles) et des métadonnées associées [Woisard et al., 2021].

Ce corpus a été collecté dans le contexte du projet INCa Carcinologic Speech Severity Index (C2SI [8]) et j'ai participé au raffinement de cette base lors du projet ANR RUGBI [9]. Il est mis à la disposition de la communauté scientifique par le biais du Groupement d'intérêt scientifique Parolothèque [10].

Tout au long de mon travail, j'ai respecté le RGPD du promoteur RC31/18/0448 [11], dont le travail sur les données devait s'effectuer :

- soit directement sur le serveur OSIRIM [12],
- soit sur copie des fichiers sur les ordinateurs cryptés des chercheurs (afin de réaliser des annotations supplémentaires, écouter les fichiers audio...).

Le projet C2SI visait à obtenir une base de données d'enregistrements vocaux français, de personnes atteintes de cancers ORL, visant à valider les index perceptifs de troubles de la parole tels que la sévérité et l'intelligibilité. Une telle base de données est utile pour mesurer l'impact du cancer de la cavité buccale et pharyngée sur la production de la parole.

Ce corpus comporte des patients (personnes suivies médicalement) et des contrôles (personnes considérées « saines »). De ce fait, les critères d'inclusion spécifiques aux patients sont :

- avoir un cancer (taille de la tumeur suivant la classification TNM de T1 à T4 [13]) de la cavité buccale et/ou du pharynx ;

8. Projet Institut National du Cancer : https://www.irit.fr/SAMOVA/site/projects/previous/c2si/
9. Projet porté par l'IRIT : https://www.irit.fr/SAMOVA/site/projects/current/rugbi/
10. https://www.irit.fr/parolotheque/
11. Plus de détails en annexe B
12. https://osirim.irit.fr/
13. Avec les tailles T1 et T2 considérées comme des tumeurs de petite taille, tandis que les tailles T3 et T4 sont considérées comme de grandes tailles. Pour plus de détails : https://www.onco-hdf.fr/app/uploads/2019/11/R%C3%A9f%C3%A9rentiel-r%C3%A9gional-VADS-VF.pdf.

— avoir été traité par chirurgie et/ou radiothérapie et/ou chimiothérapie ;
— avoir plus de six mois après la fin du traitement pour assurer la stabilité du trouble de la parole, qu'il soit audible ou non.

Tandis que les critères de non-inclusion des contrôles et des patients sont :
— présenter une autre source de troubles de la parole (entre autres, le bégaiement) ;
— présenter des problèmes cognitifs ou visuels incompatibles avec certaines tâches à effectuer par les participants.

En plus des enregistrements effectués, les patients ont répondu à un questionnaire de qualité de vie. Les résultats de ce questionnaire étant en dehors de la problématique centrale de ma thèse, je n'ai pas traité les données de ce questionnaire.

Ce corpus a aussi permis d'évaluer plusieurs niveaux d'intelligibilité/sévérité, mais également de compréhensibilité des fonctions vocales. Pour arriver à cette diversité de résultats, il est constitué de plusieurs tâches orales réalisées par les participants. Nous détaillerons ces différentes tâches dans la première sous-partie de ce chapitre.

L'étude de ce corpus se découpe en deux sections. Dans la première, nous explorons les données de la vérité terrain. Tandis que dans la seconde, nous effectuons une analyse conjointe des enregistrements sonores et de la vérité terrain.

3.1 Description des tâches

3.1.1 Tâches liées à la parole

Tâche du /a/ tenu

Cette tâche consiste en la production de trois /a/ tenu durant quelques secondes. Une voyelle tenue donne des informations sur le niveau de la voix, le temps de phonation, la stabilité, le contenu des harmoniques, le bruit, les segments non voisés, etc. Ceci étant une condition minimale pour une production vocale correcte.

Tâche de pseudo-mots

Cette tâche consiste en la prononciation de 52 pseudo-mots parmi 90 000 pseudo-mots générés automatiquement. Le locuteur était placé devant un écran et les pseudo-mots s'affichaient automatiquement tandis qu'une version sonore était diffusée de manière synchrone. La particularité de ces pseudo-mots est qu'ils ne font pas partie des mots de la langue française (tel que spofo, gorquin ou soublou).

Tâche de vérification de phrases

Cette tâche consiste en la lecture de 50 phrases sélectionnées dans une liste de 300 phrases vraies ou fausses. Une phrase « Vraie » est une phrase faisant une affirmation dont le sens est vrai (exemple : « Un éléphant est un grand animal. »), tandis qu'une phrase « Fausse » est une phrase dont le sens est faux (exemple : « Un éléphant est un animal minuscule. »).

Tâche de lecture

Cette tâche consiste en la lecture du premier paragraphe de « La chèvre de M. Seguin », un conte d'Alphonse Daudet. Ce texte présente une assez bonne couverture de l'inventaire phonétique français. Il est également bien connu et répandu en phonétique clinique en France [Ghio et al., 2012].

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

3.1.2 Tâches prosodiques

Tâche de modalité

Cette tâche consiste en la production de dix phrases identiques avec trois modalités différentes : assertion, question et exclamation (« Tu manges du jambon ./ ?/ ! »).

Tâche de focus

Cette tâche consiste en la production de vingt phrases pour lesquelles les participants doivent marquer l'accent en soulignant l'information importante de l'énoncé à l'aide de seuls indices prosodiques. L'information importante étant au préalable notifiée au locuteur vu que les phrases peuvent avoir plusieurs informations importantes (par exemple : « Avez-vous vu un chien ou un éléphant dans ce jardin ? »).

Tâche de désambiguïsation syntaxique

Chaque locuteur a effectué une lecture de 13 phrases avec deux conditions syntaxiques (portée étroite ou large de l'adjectif). Les conditions syntaxiques permettent de désambiguïser des écritures similaires. Ainsi, dans la phrase « les chevaux et les poneys blancs », l'adjectif « blancs » peut s'appliquer soit au deuxième nom seulement (« poneys », il s'agit d'une portée étroite), soit aux deux noms (« chevaux et poneys », il s'agit d'une portée large).

Tâche de description

Le participant décrit oralement une image qu'il a au préalable choisie (toutes les images comportant un bateau avec la mer). L'objectif étant qu'un examinateur puisse redessiner cette image grâce à la description du participant.

Tâche de parole spontanée

Le participant devait (pendant au moins trois minutes) donner son avis sur le questionnaire qu'il avait rempli (questionnaire sur la qualité de vie).

Maintenant que nous avons décrit le contexte et les objectifs de ce corpus, regardons en détail les données que nous avons à disposition.

3.2 Analyse de la vérité terrain

Le corpus cancer est distribué sous forme d'un répertoire contenant un fichier qui regroupe la vérité terrain (score des experts, informations des participants et des réponses à des questionnaires de qualité de vie) ainsi que des dossiers contenant les enregistrements effectués pour chaque tâche du corpus.

Nous allons analyser la vérité terrain suivant deux points de vue :

— les informations concernant les participants,

— les scores donnés par le jury d'experts,

3.2.1 Informations sur les participants

Détails communs aux contrôles et aux patients

Le corpus cancer est composé d'enregistrements audio de 128 participants (patients et contrôles) et des métadonnées associées. Cependant, nous avons uniquement la vérité terrain pour 127 participants. Il s'agit de 87 patients et de 40 contrôles. Ils sont distribués en hommes et femmes avec un léger déséquilibre en faveur du nombre d'hommes d'environ 9% (voir la figure 3.1 pour plus de détails). Ce déséquilibre est présent sur les patients (avec environ 15% d'hommes en plus sur le groupe de patients) et c'est légèrement l'inverse pour les contrôles (avec environ 10% de femmes en plus sur le groupe de contrôles).

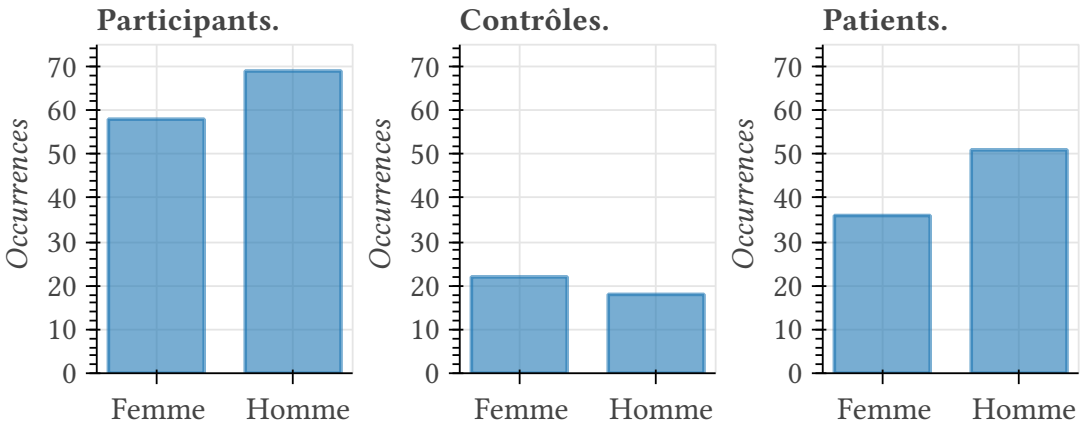


FIGURE 3.1 – Distribution du sexe des participants, avec distinction contrôles et patients.

Tous les patients, ainsi que 26 contrôles, ont été enregistrés à l'Oncopole de Toulouse. Les 14 derniers contrôles ont été enregistrés au Laboratoire de Parole et de Langage (LPL) d'Aix-en-Provence. Cela se traduit sur la distribution des lieux de résidences des participants (illustré dans la figure 3.2) avec une majorité de participants résidents de Haute-Garonne tout en ayant des participants de 13 autres départements.

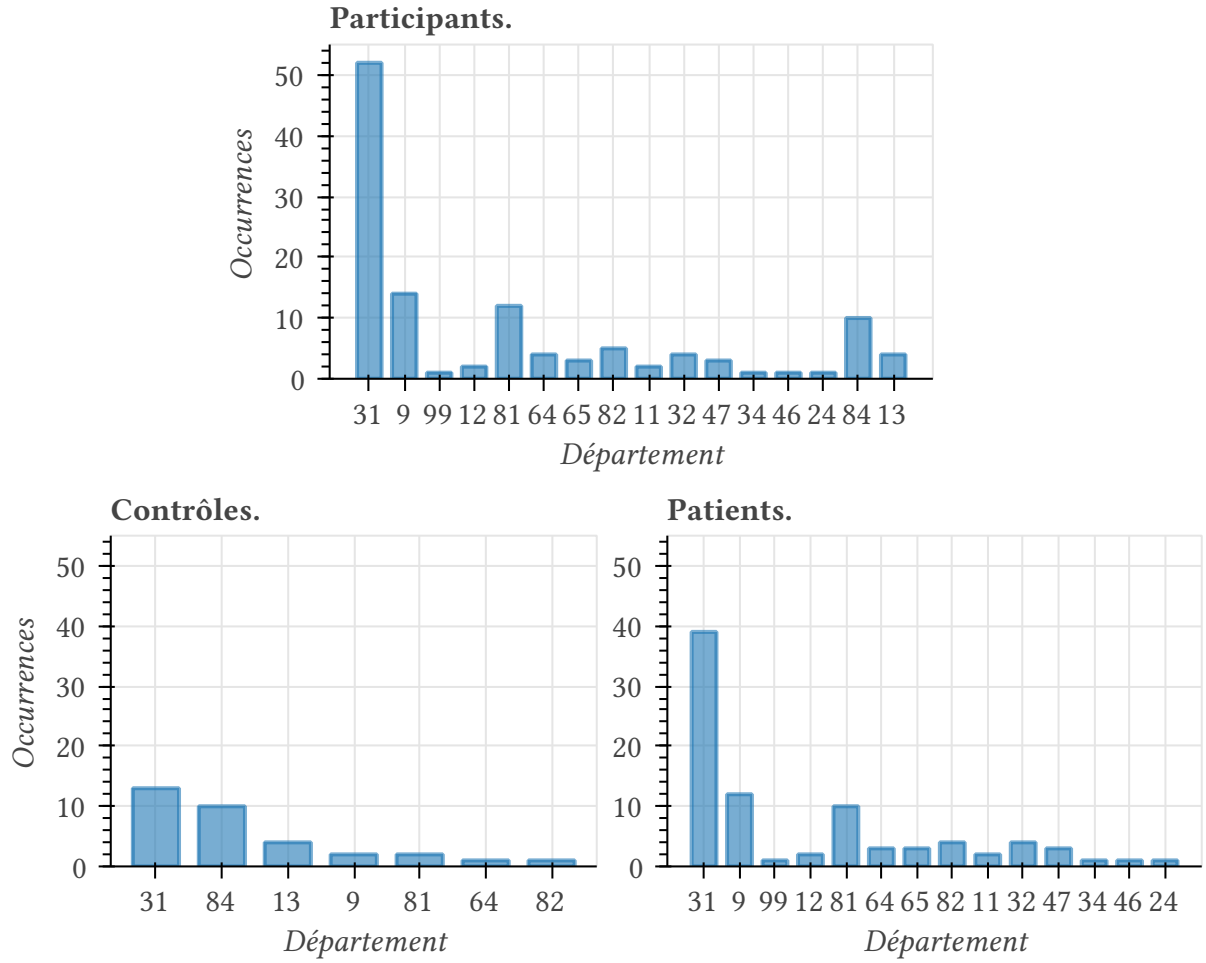


FIGURE 3.2 – Distribution des lieux de résidence des participants, avec distinction contrôles et patients.

Pour finir cette section sur les informations communes entre les patients et les contrôles de ce corpus, regardons la distribution des âges des participants (voir figure 3.3). Nous pouvons remarquer que la tranche d'âge des participants se situe majoritairement de 50 à 80 ans et que les moins de 45 ans sont surtout représentés par les contrôles.

Certains participants ont participé plusieurs fois, une participation sera appelée session (avec un maximum de deux sessions par participant, ces sessions étant espacées de plusieurs jours). Ceci permet d'ajouter une dimension temporelle sur certaines participations. Au total, il y a 139 sessions de disponibles.

Détails sur les données spécifiques aux patients

Focalisons-nous maintenant sur les informations disponibles uniquement pour les patients (car dépendant de la maladie). Un résumé des informations binaires (avec des valeurs vraies ou fausses) se trouve en figure 3.4. Nous pouvons constater qu'à part pour la chimiothérapie (avec

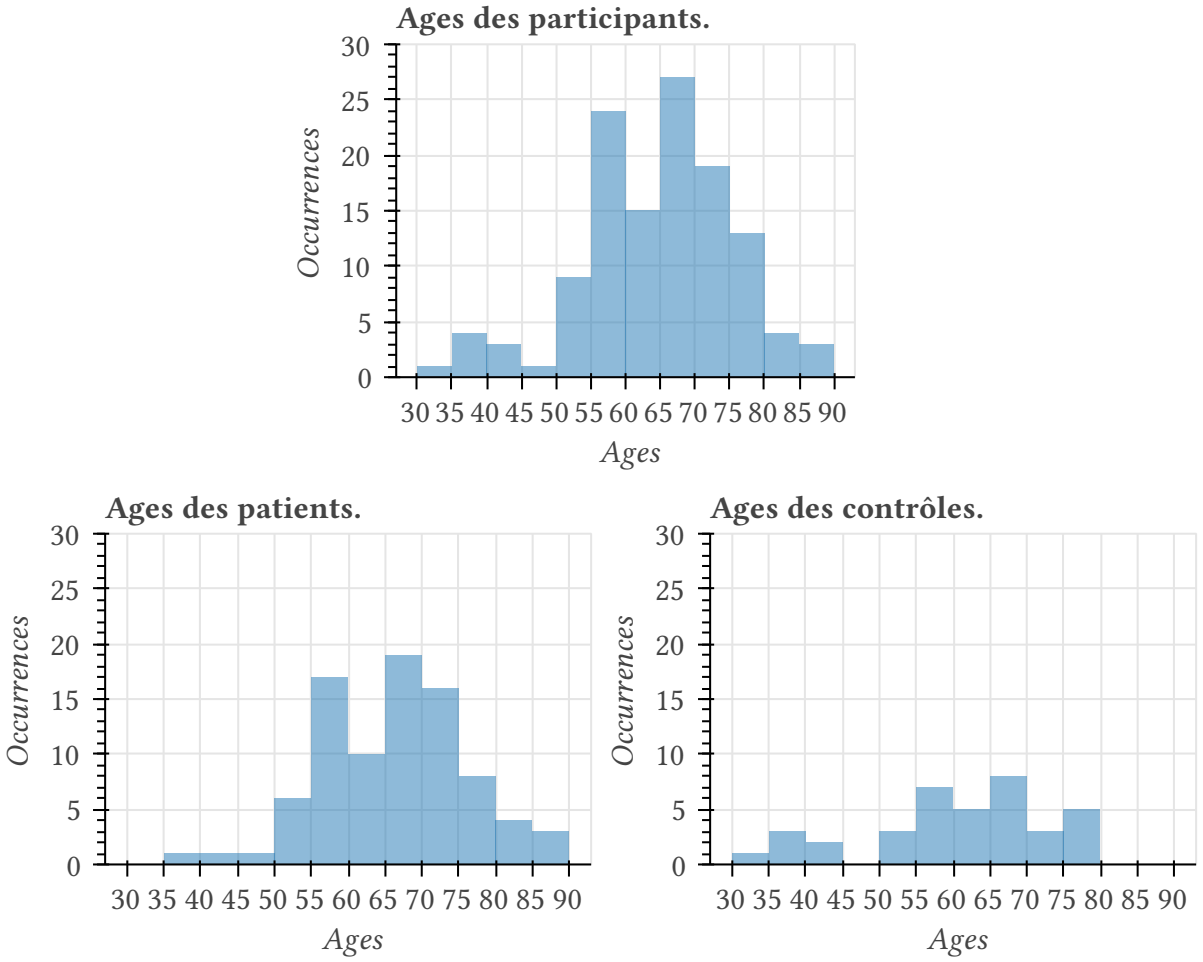


FIGURE 3.3 – Distribution des âges des participants, avec distinction contrôles et patients.

48 patients soit environ 55% des patients), les caractéristiques des patients sont déséquilibrés dans un objectif de modélisation automatique. Ceci s'explique, car les patients avec une petite tumeur peuvent avoir soit de la chirurgie seule, soit de la radiothérapie seule. C'est lorsque la maladie est plus volumineuse et avec un pronostic plus alarmant que les traitements sont combinés. On a majoritairement des patients ayant eu une chirurgie (avec 73 patients soit environ 84% des patients), ayant eu une radiothérapie (avec 82 patients, soit environ 94% des patients), n'ayant pas eu de récidive (avec 62 patients, soit environ 71% des patients) et ayant eu une reconstruction (avec 58 patients, soit environ 67% des patients).

Parmi les chirurgies suivies par les patients, nous retrouvons des glossectomies (pour 24 patients, environ 33% des patients ayant suivi une chirurgie), des mandibulectomies (pour 24 patients, soit environ 33% des patients ayant suivi une chirurgie), des pelvectomies (pour 27 patients, soit environ 37% des patients ayant suivi une chirurgie) et des oropharyngectomies (pour 26 patients, soit environ 36% des patients ayant suivi une chirurgie). Les autres chirurgies sont moins représentées (avec moins de cinq patients, soit un peu moins de 7% des patients ayant suivi une chirurgie). Pour plus de détails, voir la figure 3.5.

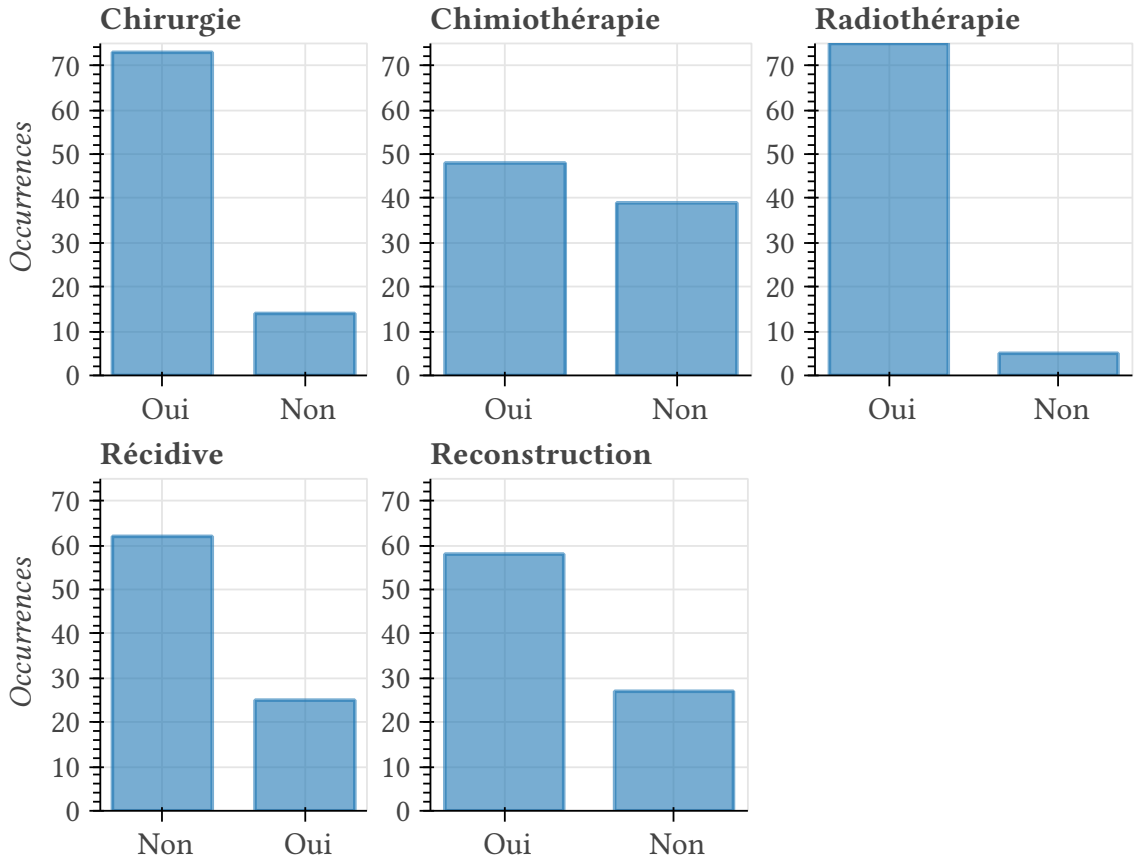


FIGURE 3.4 – Distribution de valeurs binaires concernant les patients.

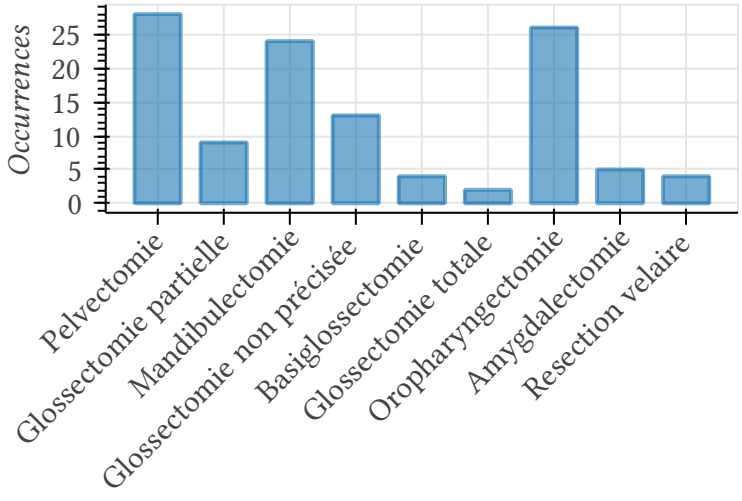


FIGURE 3.5 – Distribution du type de chirurgies suivies par les patients.

Nous avons également à disposition la taille « T » de la tumeur des patients selon le critère TNM. Le critère T de la classification TNM de la tumeur [Brierley, 2016] est une valeur allant

de 0 pour l'absence de tumeur à 4 pour une tumeur imposante et envahissante. La figure 3.6 nous montre la distribution de ces valeurs pour tous les participants (sachant que tous les contrôles ont une taille « T » de 0). Nous pouvons constater qu'environ 74% des patients ont soit une taille « T » de 2, soit une tumeur de taille « T » de 4. Les tailles « T » de 1 et 3 étant sous représentés dans le corpus cancer.

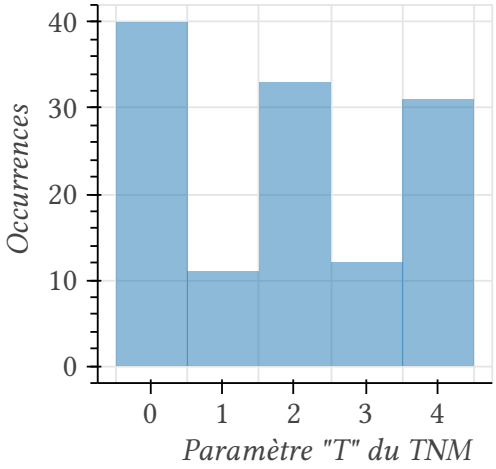


FIGURE 3.6 – Distribution des tailles « T » (selon la classification TNM) des participants. La taille 0 correspond aux contrôles dénués de tumeurs buccales.

Enfin, nous disposons de la localisation de la tumeur pour chaque patient. La figure 3.7 illustre la distribution de ces localisations (au nombre de huit sur ce corpus). Ainsi, nous nous apercevons que les tumeurs localisées sur l'amygdale (environ 30% des patients), la langue (base de la langue et langue, avec environ 25% des patients) et le plancher (environ 18% des patients) sont les plus représentées sur ce corpus. Tandis que des tumeurs localisées sur la voile (environ 5% des patients), la mandibule (avec environ 6% des patients) et la rétromolaire (environ 7% des patients) sont sous-représentées sur ce corpus.

Maintenant que nous avons vu les informations relatives aux participants, analysons les scores attribués aux participants.

3.2.2 Scores attribués aux participants

L'évaluation de la qualité de production des participants s'est effectuée au travers de différentes mesures perceptives. Néanmoins, toutes les tâches n'ont pas eu d'analyse perceptive. Ainsi, les tâches de maintien d'un /a/, de vérification des phrases et de parole spontanée manquent d'évaluations perceptives. Pour cette raison, j'ai écarté les tâches de vérification de phrase et de parole spontanée des tâches utilisables pour ma thèse. J'ai néanmoins gardé la tâche du /a/ tenu, puisque la tâche est relativement simple et permet de tester quelques idées, même s'il n'y a pas d'évaluation perceptive pour cette tâche. Pour la suite, nous analysons les scores perceptifs par tâche.

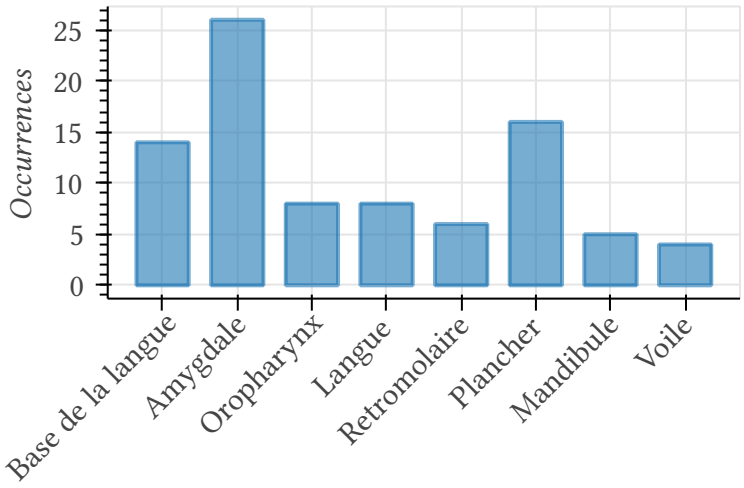


FIGURE 3.7 – Distribution de la localisation des tumeurs des patients.

Tâche de pseudo-mots

Pour la tâche de pseudo-mots, nous disposons seulement du trait d'écart moyen par phonème. Cette mesure effectuée par un annotateur consiste à évaluer la disparité moyenne entre les phonèmes perçus et ceux attendus. Pour cette tâche, les valeurs possibles vont de 0 (absence d'altération) à 5 (aucun phonème reconnu n'est correct).

Sur la figure 3.8, nous retrouvons la distribution des scores obtenus par les participants. Parmi ces scores, la valeur minimale est de 0,19, la valeur maximale est de 4,07 et l'écart-type des différents scores obtenus est ce 0,65. Les scores au-dessus de trois sont rares : seulement un cas. Par ailleurs, les scores obtenus sont déséquilibrés sur les tranches de valeurs allant de 0 à 3. De plus, sur la figure 3.8, nous remarquons que les tumeurs des patients ont tendance à influencer négativement les scores obtenus par le groupe de patients par rapport au groupe contrôle (cf. boîte à moustache entre le score obtenu par les participants et le paramètre « T » du TNM).

Nous avons rejeté cette tâche pour la suite de ma thèse du fait que cette tâche contient des pseudo-mots (différents à chaque fois), qu'il n'existe pas de mesure perceptive du concept d'intelligibilité (ou de sévérité), et que les scores sont très mal distribués.

Tâche de Lecture

Cette tâche a été évaluée par un jury de cinq experts. Ces experts devaient donner des scores de sévérité du trouble de la parole (échelonnés de 0 à 10 : de très atteint à pas atteint), mais également donner des scores d'intelligibilité de production (échelonnés de 0 à 10 : de non intelligible à très intelligible). Cependant, des données sont manquantes dans la vérité terrain dont nous disposons, comme illustré dans le tableau 3.1. La valeur moyenne peut se retrouver dans la majorité des cas grâce aux scores individuels des membres du jury. En effet, bien qu'il y ait cinq experts pour 48 participants, la majorité des scores ont été créés avec trois

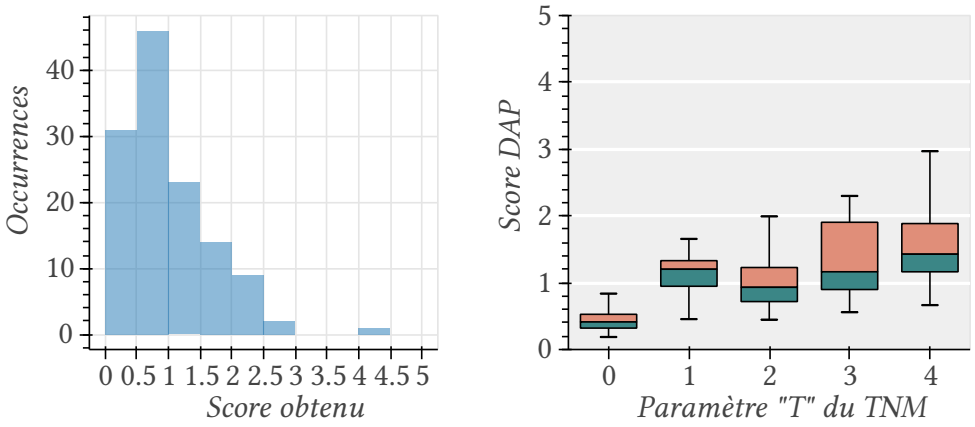


FIGURE 3.8 – Distribution des scores de la tâche de pseudo-mots : traits d'écart moyen par phonème (gauche) et boîte à moustache en fonction du paramètre « T » du TNM.

experts seulement. Nous remarquons également que le nombre de scores moyens [14] (115) est inférieur aux nombres de sessions effectués (139) : nous avons un score pour 83% des sessions d'enregistrements.

TABLE 3.1 – Nombre de scores disponibles pour la tâche de lecture.

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Score moyen
Sévérité	48	113	48	114	114	115
Intelligibilité						

Par la suite, nous utiliserons uniquement les scores moyens que nous appellerons score perceptif de sévérité et score perceptif d'intelligibilité. Les distributions de ces scores sont présentes sur la figure 3.9. Celle de l'intelligibilité est plus homogène que celle de la sévérité.

Il y a également plus de tranches de valeurs pour le score perceptif de sévérité que pour celui d'intelligibilité. Ceci était attendu, vu que la sévérité mesure des altérations non décelables par l'intelligibilité telles que des paramètres temporels et/ou prosodiques [Auzou et al., 2007]. Cela se traduit par une valeur minimale des scores d'intelligibilité de 3,5 contre une valeur minimale de 1,03 pour le score perceptif de sévérité. Par ailleurs, l'écart-type de l'ensemble des scores d'intelligibilité est de 1,27 tandis que sur le score sévérité, il est de 2,31. Sachant que ces deux scores sont fortement corrélés (avec une corrélation de Spearman de 0,94, voir répartition sur la figure 3.10), il est préférable d'utiliser le score perceptif de sévérité comme score de référence.

Sur la figure 3.11 nous retrouvons les boîtes à moustaches des scores perceptifs avec la taille « T » (selon la classification TNM) des tumeurs des participants (rappel, 0 étant les contrôles).

14. Le score moyen représente la moyenne entre les scores disponibles des experts. Il est à noter que certains scores d'experts étant manquants, nous ne pouvons pas toujours calculer ce score moyen. Néanmoins, nous utiliserons ce score lorsque nous parlons de score moyen par la suite comme il nous permet d'utiliser quelques fichiers supplémentaires dans nos expériences et d'avoir un score moyen au plus proche d'un jury de 5 experts pour chaque fichier.

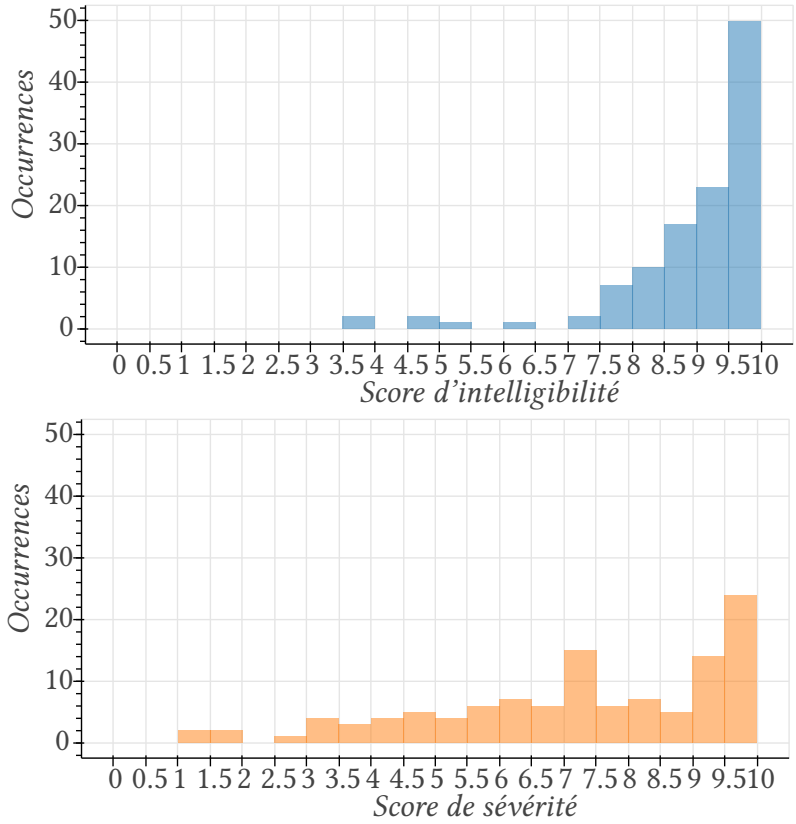


FIGURE 3.9 – Distribution des scores de sévérité et d'intelligibilité sur la tache de lecture.

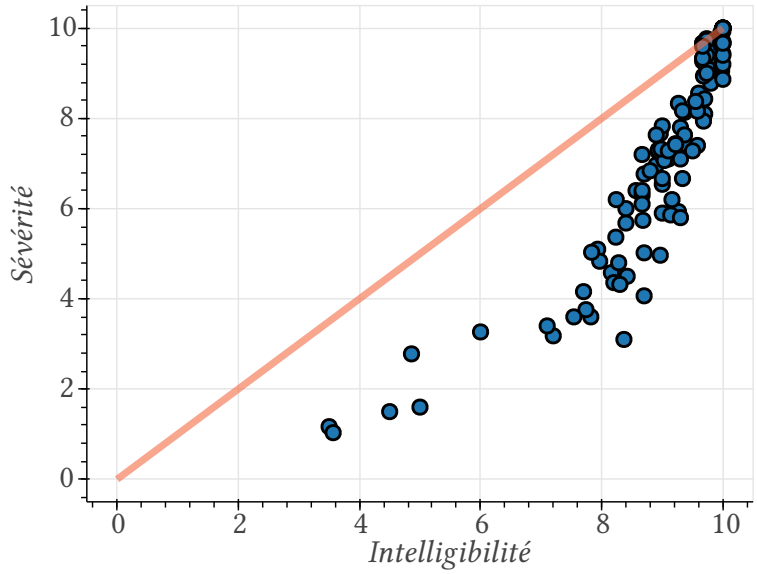


FIGURE 3.10 – Comparaison des scores d'intelligibilité et de sévérité obtenus pour la tâche de lecture.

Il est à noter que le score perceptif de sévérité possède la corrélation de Spearman la plus élevée avec la taille de la tumeur TNM (de -0,65 contre -0,63 pour l'intelligibilité).

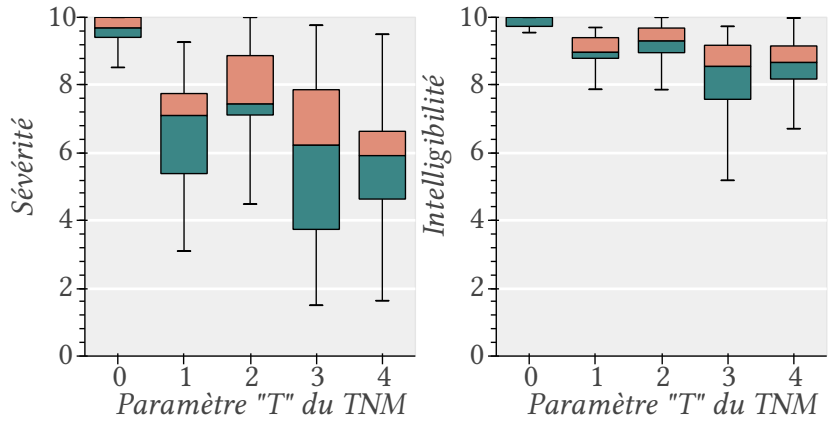


FIGURE 3.11 – Boîtes à moustaches des tailles « T » (selon la classification TNM) de tumeurs et des scores données sur la tâche de lecture.

Tâche de description

Pour cette tâche, comme pour la tâche de lecture, les experts du jury (ici au nombre de six) devaient évaluer l'intelligibilité et la sévérité des patients. Mais, avant cela, ils ont dû évaluer quatre autres critères qui sont des mesures d'altération de :

- la résonance,
- la prosodie,
- la voix,
- la prononciation phonémique.

Ces critères sont échelonnés de 0 (pas d'altération) à 3 (grande altération). Comme pour la tâche de lecture, des données sont manquantes pour certains scores individuels d'experts (voir table 3.2). De surcroît, comme pour la tâche de lecture le nombre de scores moyens [15] est inférieur aux nombres de sessions effectués (nous avons aussi un score pour environ 83% des sessions d'enregistrements). Le nombre de mesures de perception pour cette tâche étant conséquent, nous avons séparé notre analyse en deux en fonction des échelles de valeurs des mesures perceptives.

TABLE 3.2 – Nombre de scores disponibles pour la tâche de description.

	Expert 1	Expert 2	Expert 3	Expert 4	Expert 5	Expert 6	Score moyen
Sévérité	110	110	111	104	112	111	115
Intelligibilité							
Résonance	110	108	109	102	110	109	113
Voix							
Prosodie							
Phonémique							

15. Tout comme pour la tâche de lecture, le score moyen représente la moyenne entre les scores disponibles des experts.

Le nombre de mesures de perception pour cette tâche étant conséquent, nous nous sommes concentrés sur les scores moyens. Ainsi, nous avons réalisé une table de compilation des corrélations de Spearman entre ces différentes mesures (voir table 3.3). Nous pouvons constater que les corrélations sont globalement fortes, excepté pour l'altération de la voix. Les scores de sévérité et d'intelligibilité sont les scores les plus corrélés avec l'altération phonémique. Au vu du protocole d'évaluation du jury (l'évaluation de l'altération phonémique a été effectuée avant de donner les scores de sévérité et d'intelligibilité), nous pouvons penser que l'évaluation du score phonémique a influencé les experts sur l'évaluation du score perceptif de sévérité et d'intelligibilité.

TABLE 3.3 – Corrélations entre les différents scores de perception obtenus sur la tâche de description.

Score	Intelligibilité	Sévérité	Voix	Résonance	Prosodie	Phonémique
Intelligibilité		0,92	-0,63	-0,86	-0,81	-0,93
Sévérité	0,92		-0,68	-0,88	-0,77	-0,92
Voix	-0,63	-0,68		0,51	0,68	0,55
Résonance	-0,86	-0,88	0,51		0,71	0,83
Prosodie	-0,81	-0,77	0,68	0,71		0,75
Phonémique	-0,93	-0,92	0,55	0,83	0,75	

La distribution des scores d'altération est représentée dans la figure 3.12 et les boîtes à moustaches de ces scores avec la taille « T » de la tumeur se trouvent sur la figure 3.13.

Nous pouvons constater que le score phonémique est le mieux distribué. Ceci permettrait de modéliser plus facilement ce score. Nous pouvons également observer que les scores d'altération de résonances et phonémique sont plus corrélés à la taille « T » de la tumeur (0,64 et 0,70 respectivement) que les scores de voix et de prosodie (0,39 et 0,53 respectivement). Néanmoins, comme le score d'altération phonémique est très fortement corrélé aux scores de sévérité et d'intelligibilité, nous allons l'écarter (ainsi que les trois autres scores d'altérations) des scores sur lesquels nous allons nous focaliser.

En figure 3.14 nous avons les distributions des scores de sévérité et d'intelligibilité. La distribution du score perceptif de sévérité reste la plus hétérogène avec un score minimal de 0,58 et un écart-type de 2,57. Sur cette tâche, la distribution du score d'intelligibilité est plus hétérogène que sur la tâche de lecture.

Les scores d'intelligibilité et de sévérité sont fortement corrélés (voir figure 3.15) : ceci nous laisse penser qu'une simple fonction linéaire ferait le lien entre ces deux scores.

Les scores d'intelligibilité et de sévérité sont comme le score d'altération phonémique corrélés à la taille « T » de la tumeur avec une corrélation de -0,65 pour l'intelligibilité et de -0,67 pour la sévérité. La figure 3.16 illustre ce résultat et nous indique que l'intelligibilité est sur cette tâche autant intéressante que le score perceptif de sévérité.

Tâches de prosodies

Les juges devaient pour la tâche de focus repérer l'élément mis en avant, tandis que pour la tâche de modalité, ils devaient repérer le type de phrase produite pour enfin repérer si seul

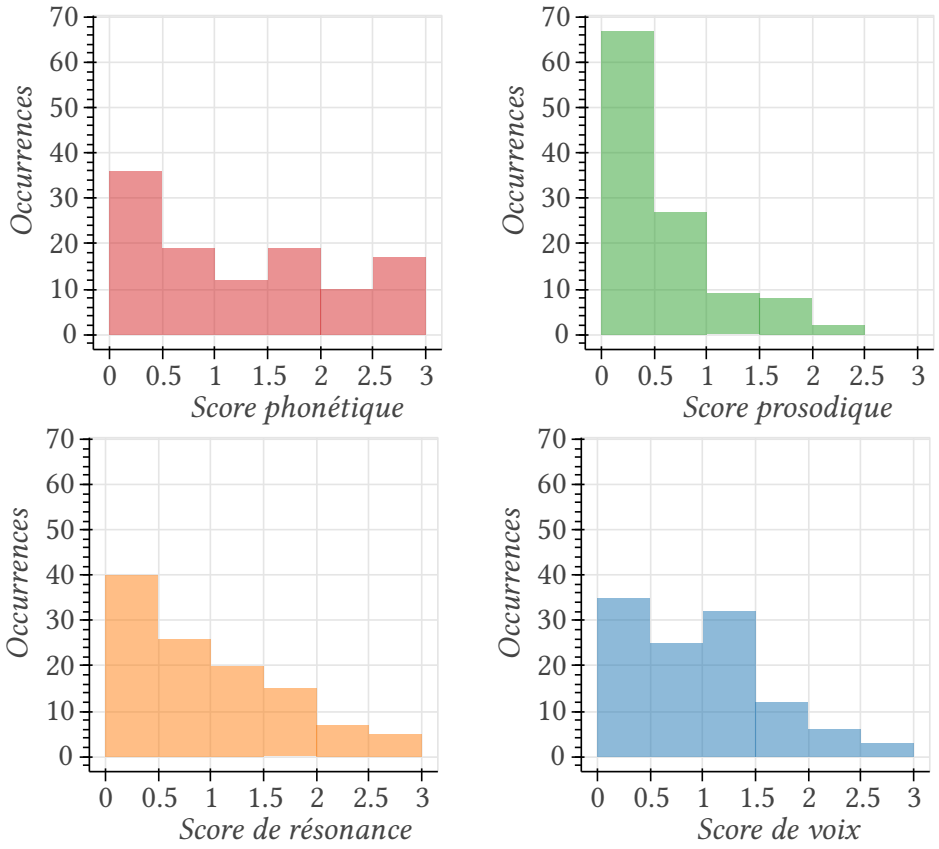


FIGURE 3.12 – Distribution des scores d'altération de résonance, prosodique, de prononciation phonémique et de voix sur la tâche de description. Plus le score est proche de zéro, plus la parole prononcée est compréhensible.

un nom ou au contraire tout un groupe nominal étaient concernés par le qualificatif pour la tâche de désambiguïsation syntaxique. Ainsi, ces trois tâches ont permis de créer trois scores prosodiques échelonnés de 0 à 3. Plus le score est faible, moins la prosodie est correctement retrouvée par les auditeurs.

Sur les figures 3.17, 3.18 et 3.19 nous pouvons retrouver les distributions de ces trois scores ainsi que les boîtes à moustaches de ces scores avec la taille « T » de la tumeur des participants. Ainsi, ces trois scores sont très peu corrélés de la taille « T » de la tumeur (-0,19 pour le score perceptif de focus, -0,29 pour le score perceptif de modalité et -0,19 pour le score perceptif de désambiguïsation syntaxique). Ceci peut nous laisser penser que ces mesures sont décorrélées des scores d'intelligibilité et de sévérité.

Ces trois scores sont les scores perceptifs dont la distribution des valeurs est la plus homogène parmi toutes les tâches du corpus cancer. Le score perceptif de modalité a un écart-type de 0,42, une valeur minimale de 0,62 et une valeur maximale de 2,85 ; le score perceptif de focus a un écart-type de 0,37, une valeur minimale de 1,21 et une valeur maximale de 2,73 ; le score perceptif de désambiguïsation syntaxique a un écart-type de 0,37, une valeur minimale de 1,35 et une valeur maximale de 2,71. Ainsi, avec des écarts-types les plus faibles des scores

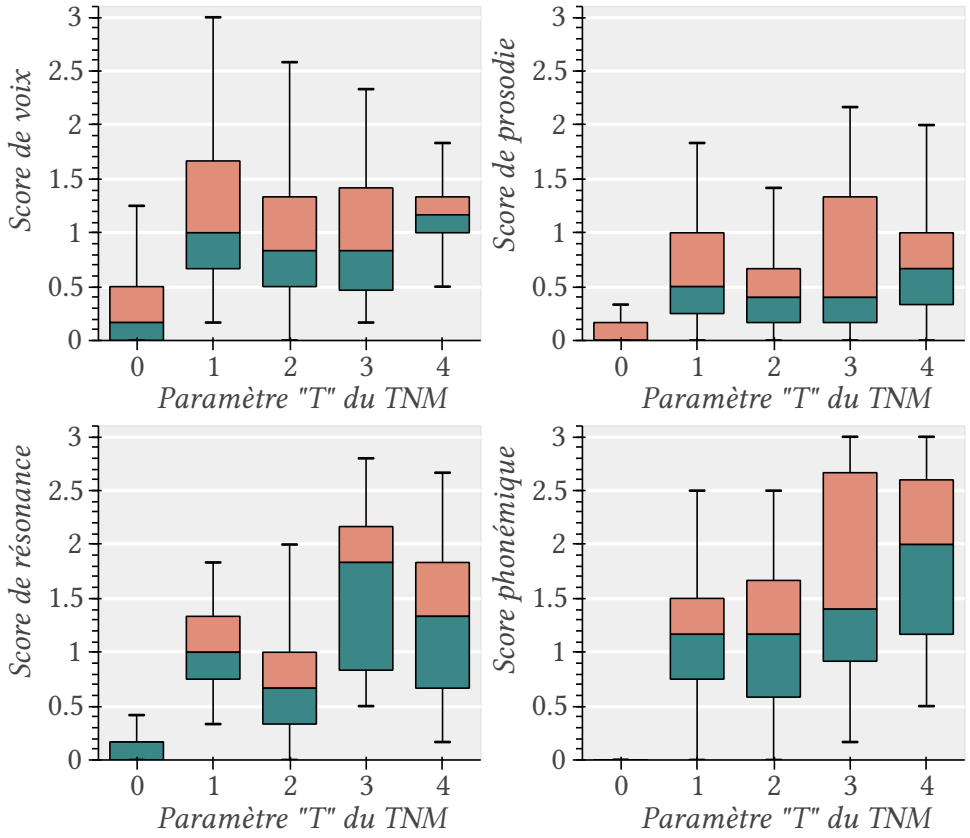


FIGURE 3.13 – Boîtes à moustaches des tailles « T » (selon la classification TNM) de tumeurs et scores d'altération de voix, de résonance, de prosodie et de prononciation phonémique sur la tâche de lecture.

perceptifs et des valeurs minimum et maximum peu distantes, nous disposons d'une distribution très homogène. Ceci rend potentiellement plus difficile l'apprentissage de ce concept par un modèle. Par conséquent, nous éviterons ces mesures et les données associées à ces tâches pour la suite de la thèse. En effet, la nature des données les rend plus complexes à traiter que les données des autres tâches.

Suite à cette analyse des annotations, voici la liste des tâches retenues pour la suite de la thèse :

— le maintien de la voyelle /a/,
— la description d'image,
— la lecture.

Regardons maintenant plus en détail les données dont nous disposons pour ces trois tâches.

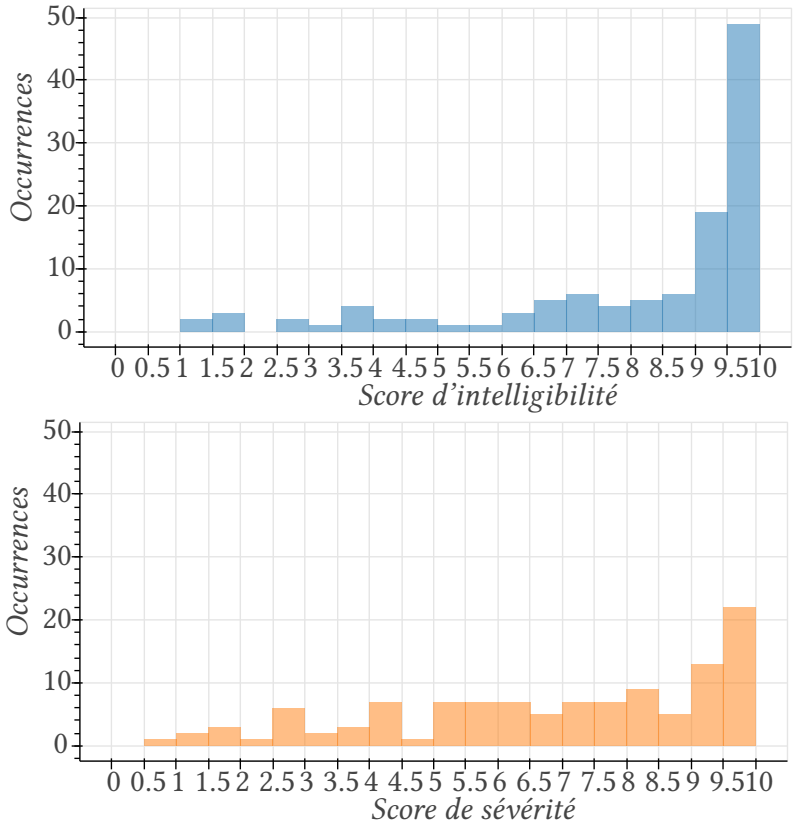


FIGURE 3.14 – Distribution des scores de sévérité et d'intelligibilité sur la tâche de description.

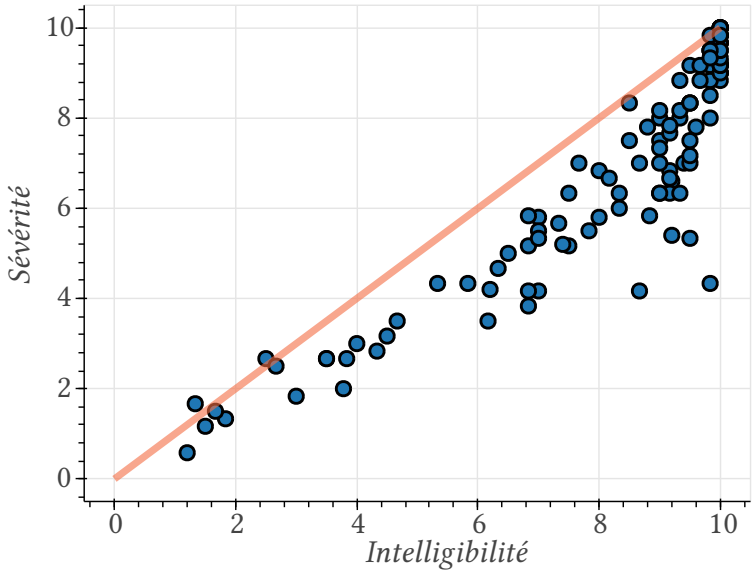


FIGURE 3.15 – Comparaison des scores d'intelligibilité et de sévérité obtenus pour la tâche de description. En rouge, nous avons la fonction $y = x$.

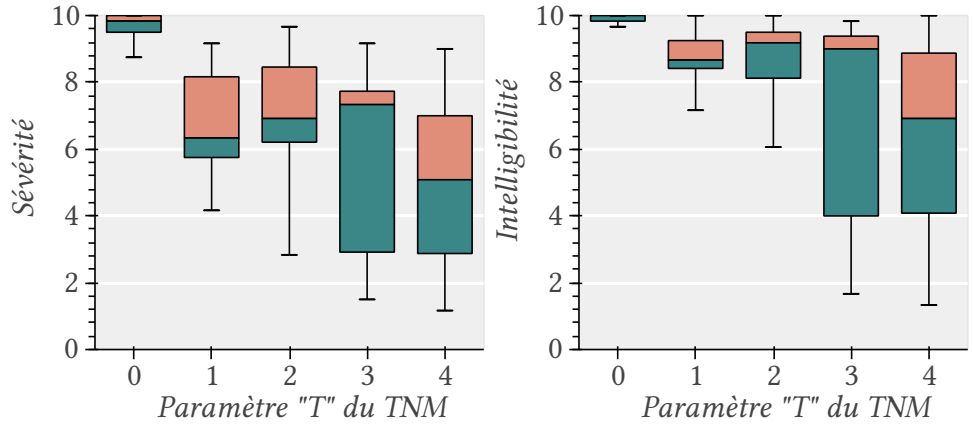


FIGURE 3.16 – Boîtes à moustaches des tailles « T » (selon la classification TNM) de tumeurs et des scores données sur la tâche de description.

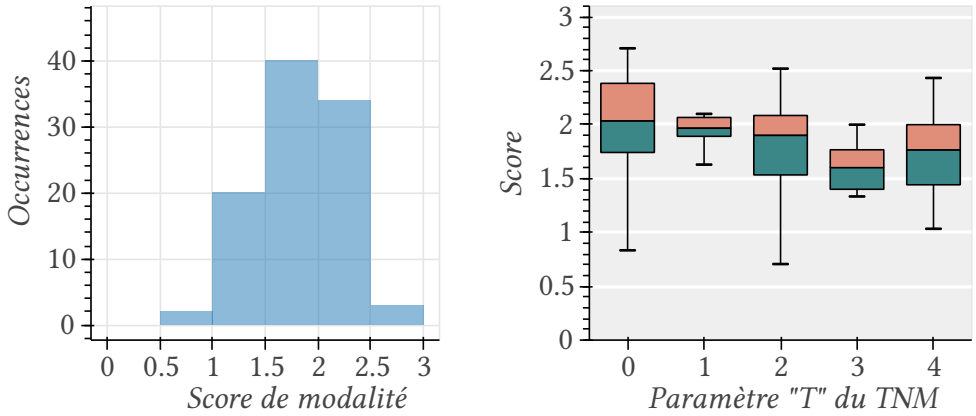


FIGURE 3.17 – Distribution du score perceptif de la tâche de modalité attribué aux participants (gauche) et boîte à moustache de ce score avec la taille « T » des tumeurs des participants (droite).

3.3 Analyse conjointe de la vérité terrain et des données du corpus

La tâche de /a/ tenu possède le plus grand nombre de fichiers disponibles (voir table 3.4). En effet, cette tâche devait être réalisée trois fois pour chaque session afin d'observer la capacité moyenne du participant à réaliser un /a/ tenu. Nous remarquons également qu'il manque des enregistrements dans la base.

Pour la suite, nous inclurons les fichiers sans évaluation lorsque nous ferons des statistiques ne nécessitant pas de vérité terrain. Dans les deux sous-sections qui vont suivre, nous allons déterminer les durées d'enregistrements disponibles.

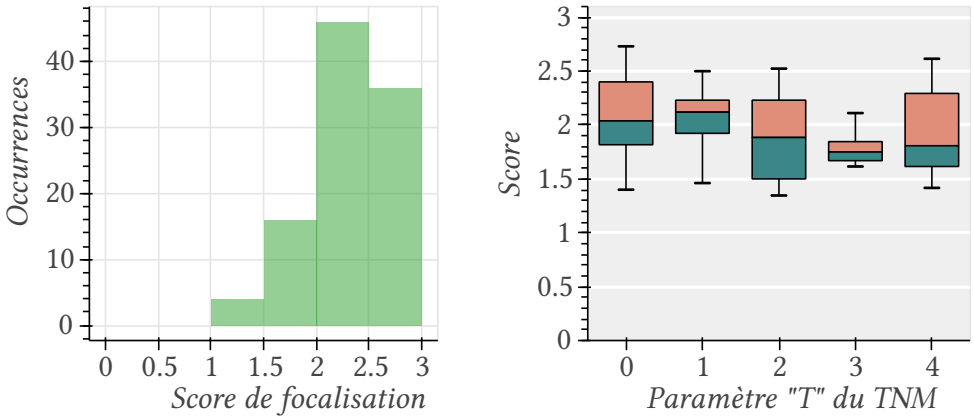


FIGURE 3.18 – Distribution du score perceptif de la tâche de focus attribué aux participants (gauche) et boîte à moustache de ce score avec la taille « T » des tumeurs des participants (droite).

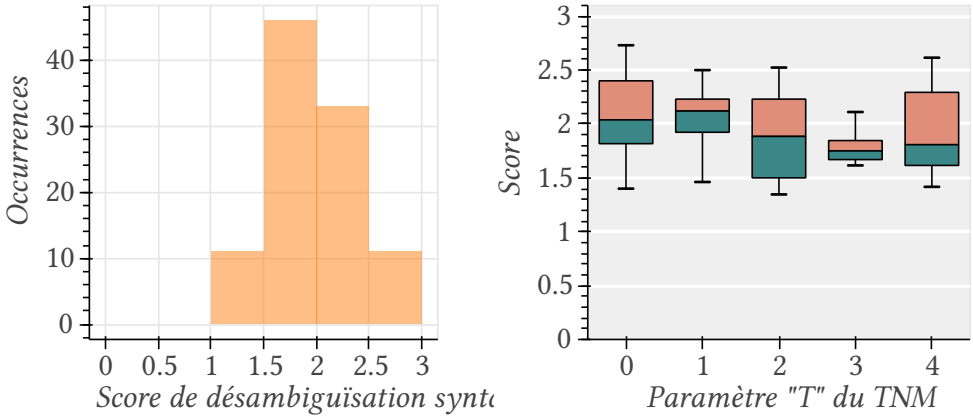


FIGURE 3.19 – Distribution du score perceptif de la tâche de désambiguïsation syntaxique attribué aux participants (gauche) et boîte à moustache de ce score avec la taille « T » des tumeurs des participants (droite).

TABLE 3.4 – Nombre de fichiers disponibles pour les tâches de /a/ tenu, de lecture et de description.

Tâche\Participants	Patients	Contrôles	Total
/a/ tenu	269	78	347
Lecture	89	25	114
Description	92	41	133

3.3.1 Durées d'enregistrement avec les données brutes

Nous utilisons les données brutes, sans les silences de début et de fin de fichier (avec la fonction *trim* de librosa [McFee et al., 2015]). L'objectif est de garder les silences inter-mots

afin d'illustrer si la maladie affecte les temps de réponse. En effet, si un patient prend plus de temps, cela peut signifier qu'il répète un ou plusieurs mots.

La quantité d'audios bruts tronqués par tâche est détaillée dans la table 3.5. Nous disposons de plus de données pour les tâches de description d'images, car il y a plus de participations et parce que la tâche est plus longue à réaliser par les participants. Ayant moins de fichiers contrôles de disponibles sur la tâche de lecture implique que nous disposons de moins de temps que la tâche de /a/ tenu.

TABLE 3.5 – Durées des données brutes pour les tâches de /a/ tenu, de lecture et de description.

Tâche\Participants	Patients	Contrôles	Total
/a/ tenu	27m 47s	12m 51s	40m 38s
Lecture	48m 03s	9m 40s	57m 44s
Description	1h 37m 21s	51m 11s	2h 28m 33s

Une illustration de la distribution de ces durées est disponible en figure 3.20. Comme attendu, le maintien du /a/ est la tâche produisant des audios de plus faible durée. Bien que la tâche de lecture permette d'obtenir des enregistrements de plus de 20 secondes, c'est la tâche de description d'images qui produit les enregistrements les plus longs.

Nous avons ensuite cherché à voir si les temps de participations étaient corrélés avec les scores de sévérité et d'intelligibilité obtenus. Pour la tâche de /a/ tenu, le plus haut score de corrélation est obtenu avec le score perceptif de sévérité de la tâche de description d'images : corrélation de Spearman de 0,21. Nous pouvons faire le même constat avec la tâche de description d'image où les durées sont faiblement corrélées aux scores d'intelligibilité (0,16 de corrélation de Spearman) et de sévérité (0,13 de corrélation de Spearman). Cependant, le constat sur la tâche de lecture est différent. En effet, les durées d'enregistrement et les scores perceptifs sont corrélés : -0,68 pour le score d'intelligibilité et -0,72 pour le score perceptif de sévérité (voir figure 3.21).

Les patients prennent plus de temps que les contrôles (tous ont un score de 10) pour effectuer la tâche. Cela peut s'expliquer par les répétitions nécessaires par les patients, des temps de pause plus longs entre les mots ou encore des temps de prononciations plus longs. Ainsi, nous avons une dimension temporelle importante sur cette tâche : ceci est certainement visible ici, car les participants devaient prononcer les mêmes phrases (contrairement à la tâche de description d'images qui est plus libre). Pour mieux comprendre ce phénomène, nous allons maintenant utiliser un détecteur d'activité vocale.

3.3.2 Durées d'enregistrement après l'application d'un détecteur d'activité vocale

Afin d'éliminer les temps d'hésitation et de pauses entre les mots, nous avons utilisé un détecteur d'activité vocale, inspiré des travaux de Moattar [Moattar and Homayounpour, 2009]. Ainsi, notre détecteur définit les segments actifs comme suit :

$$\text{VAD}(\boldsymbol{s}_t) = \left(\frac{\text{energie}(\boldsymbol{s}_t)}{\text{percentile}_k^{95}(\text{energie}(\boldsymbol{s}_k))} > \alpha \right) \wedge \left(\frac{\text{platitude spectrale}(\boldsymbol{s}_t)}{\text{percentile}_k^{95}(\text{platitude spectrale}(\boldsymbol{s}_k))} > \beta \right) \tag{3.1}$$

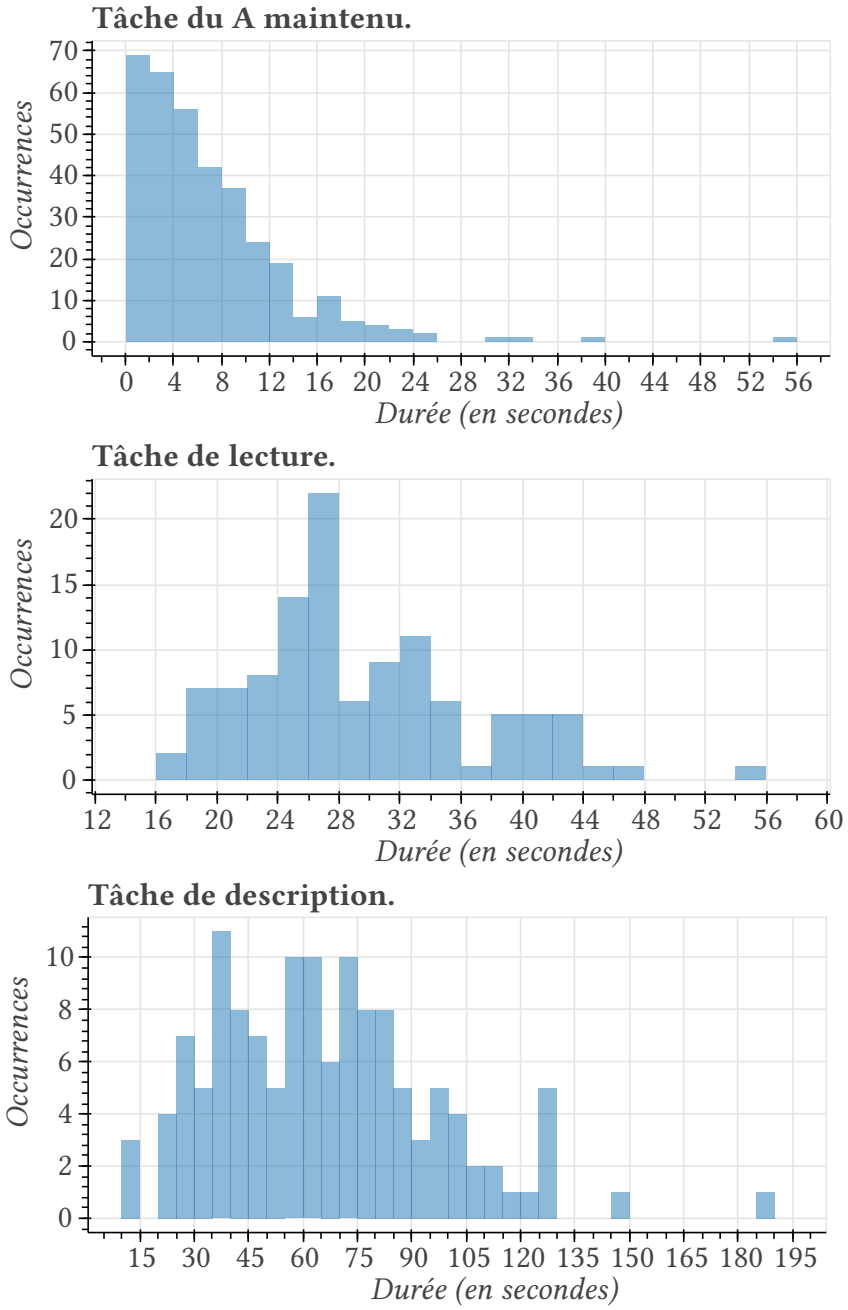


FIGURE 3.20 – Distribution des durées des fichiers bruts.

Il est appliqué sur des fenêtres glissantes de 64 ms (recouvrement de moitié). Ensuite, un segment de 5 fenêtres est considéré actif s'il contient au moins une fenêtre active. α et β ont été optimisés sur un échantillon de 20 fichiers que j'ai annotés manuellement sur le corpus cancer. Ce système obtient un F-score supérieur à 97% sur 10 autres fichiers que j'ai également annotés.

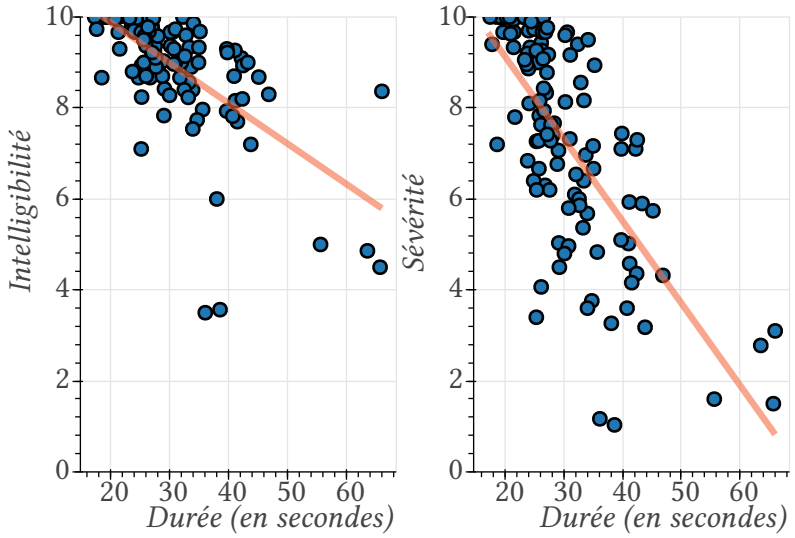


FIGURE 3.21 – Nuage de points entre les scores d'intelligibilité (gauche) et de sévérité (droite) et les durées mises pour réaliser la tâche de lecture. En rouge, est dessinée la droite de régression.

Cette fonction retourne vraie pour les segments considérés actifs avec $\boldsymbol{s}_t$ étant une fenêtre du signal $\boldsymbol{s}$ à l'instant t, α et β sont des seuils sur l'énergie et l'aplatissement spectral [Dubnov, 2004] normalisée au 95$^{\text{ième}}$ percentile de $\boldsymbol{s}$.

Notre fonction prend des décisions sur des fenêtres de 64 ms. Cette fonction est appliquée sur des fenêtres glissantes de 64 ms (avec recouvrement de moitié). Ensuite, les silences de moins de 320 ms (correspondant à cinq fenêtres) sont gardés pour s'assurer que nous supprimons uniquement les longs silences. α, β, la taille des fenêtres et le nombre de fenêtres requis pour supprimer uniquement les longs silences ont été optimisés (avec une grille de recherche sur le F-score) sur un échantillon de 20 fichiers. J'ai annoté manuellement ces fichiers provenant du corpus cancer. Ce système obtient un F-score supérieur à 97% sur 10 autres fichiers que j'ai également annotés.

Vis-à-vis de la section précédente, les durées d'enregistrement sont réduites (voir table 3.6). Les résultats sur la tâche de /a/ tenu sont pratiquement identiques, puisque seuls les silences de début et de fin sont enlevés. Il est à noter que notre détecteur d'activité est plus précis que l'outil $trim$ de la partie précédente, ce qui explique la légère différence du temps total sur les /a/ tenus. Pour la tâche de lecture, notre détecteur d'activité supprime environ un tiers du signal original. Tandis que pour la tâche de description, notre détecteur d'activité supprime environ 38% du signal original.

TABLE 3.6 – Durée des enregistrements après application d'un détecteur d'activité vocale.

Tâche\Participants	Patients	Contrôles	Total
/a/ tenu	26m 17s	12m 10s	38m 27s
Lecture	33m 08s	7m 07s	40m 16s
Description	57m 08s	36m 38s	1h 33m 46s

La corrélation entre les durées de la tâche de description et les scores de sévérité est de 0,33 (elle est de 0,34 pour l'intelligibilité). Ces valeurs sont plus importantes que sur les fichiers bruts. Nous pouvons ainsi penser que les silences enlevés correspondent en majorité à des hésitations (non corrélées aux scores d'intelligibilité et de sévérité).

Pour la tâche de lecture, la corrélation des durées avec les scores de sévérité (et d'intelligibilité) est moins importante que sur les enregistrements bruts (respectivement -0,66 pour la sévérité et -0,60 pour l'intelligibilité), voir figure 3.22. Cela nous laisse penser que ces silences ne sont pas des silences d'hésitation sur cette tâche, mais bien des silences liés aux efforts supplémentaires liés à la maladie.

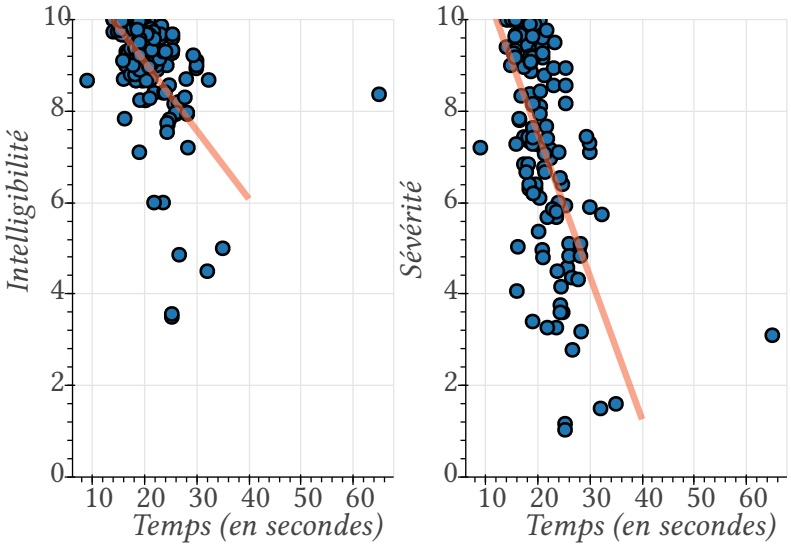


FIGURE 3.22 – Nuage de points entre les scores d'intelligibilité (gauche), de sévérité (droite) et les durées mises pour réaliser la tâche de lecture (après application d'un détecteur d'activité vocale). En rouge, est dessinée la droite de régression.

Les distributions des durées par tâche sont modifiées (voir figure 3.23).

Nous notons que les durées les plus faibles pour la tâche de lecture correspondent aux contrôles, contrairement à la tâche de description d'image où ce sont principalement les patients gravement atteints (score perceptif de sévérité faible). Ce constat nous laisse supposer que la tâche de lecture est plus intéressante pour apprendre le concept de sévérité que la tâche de description.

3.4 Discussion suite aux analyses

Nous avons à disposition de mon travail de thèse un corpus avec plusieurs tâches variées. Malgré quelques déséquilibres sur certaines caractéristiques des participants (dont leur âge), nous avons à disposition des scores avec des échelles de valeurs variées (notamment pour le score d'intelligibilité et de sévérité). Ainsi, les tâches de lecture et de description d'image sortent du lot. En effet, ce sont les deux seules tâches où les experts ont évalué l'intelligibilité des participants, mais également l'indice de sévérité. Nous avons constaté que le score perceptif de sévérité avait la répartition la moins homogène des deux scores, car l'indice de sévérité

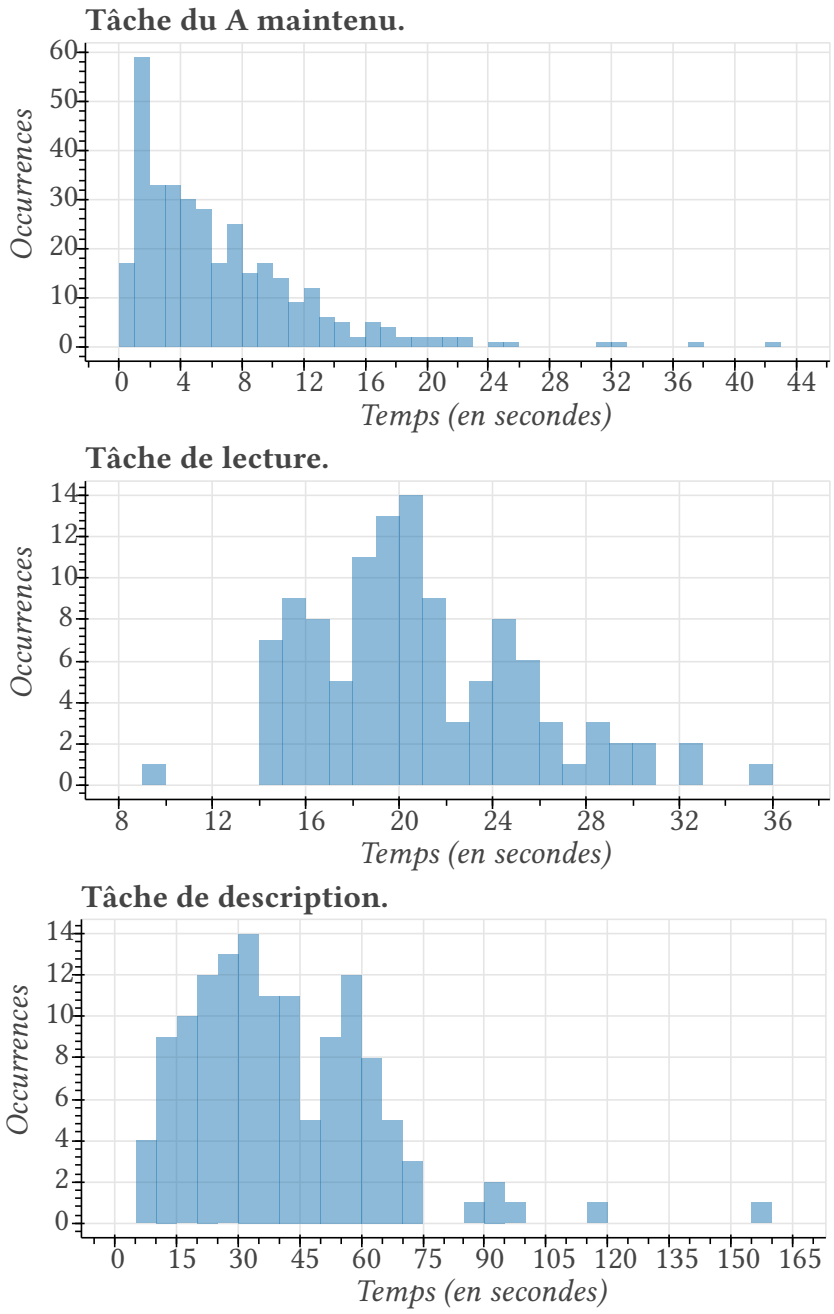


FIGURE 3.23 – Distribution des durées après utilisation d'un détecteur d'activité vocale.

décrit plus de paramètres non décelables par le score perceptif d'intelligibilité [Auzou et al., 2007]. Ceci peut jouer à notre avantage, car une répartition moins homogène peut faciliter l'apprentissage d'un modèle.

Nous avons remarqué que les durées des lectures des patients sont corrélées aux scores de sévérité et d'intelligibilité.

Contrairement à la tâche de description d'image, la tâche de lecture garantit une assez bonne couverture de l'inventaire phonétique français et nous assure de disposer d'une normalisation de la quantité de données à traiter par participant. C'est pourquoi pour la suite de mes travaux, je me suis concentré sur la tâche de lecture et la mesure de sévérité.

Conclusion de partie

Dans cette partie, dans le chapitre 1, nous avons vu l'environnement dans lequel j'ai pu évoluer : cette richesse thématique m'a permis de confronter mes idées, de plus facilement comprendre les besoins au niveau médical et d'appréhender mon travail de thèse. Par ailleurs, connaître les avancées avant la publication des papiers de mes collègues m'a aidé à affiner mes choix pour mes travaux. De plus, sachant que mes collègues partaient sur la piste de représentations x-vecteurs [Quintas et al., 2020] (version neuronale des i-vecteurs) et sur des modèles neuronaux de reconnaissance automatique de la parole [Abderrazek et al., 2020] (pour calculer la vraisemblance moyenne des mots ou des phonèmes prédits par le modèle), je me suis positionné sur une recherche complémentaire. Cela s'est traduit par l'apprentissage du concept de sévérité sur un corpus limité en quantité de données et par l'utilisation de méthodes non-supervisées pour créer une mesure semblable à l'indice de sévérité.

Pour réaliser ce travail, j'ai eu accès à un corpus cancer dont les forces sont les suivantes :

— multiples enregistrements de certains participants permettant d'évaluer la robustesse au locuteur de nos approches.
— tous les phonèmes de la langue française sont présents. Cela limite le biais de mesure des experts.
— les enregistrements ont été réalisés au sein d'un milieu hospitalier [16], ce qui permettra d'évaluer la robustesse aux bruits environnementaux de nos approches.

L'analyse fine de ce corpus (cf. chapitre 3) m'a permis, en plus d'appréhender les données, de participer à la réalisation d'un article de journal à Language Resources and Evaluation [Woisard et al., 2021]. Ainsi, ces points mettent en valeur plusieurs aspects de notre corpus, qui semble plus intéressant que ceux recensés en chapitre 2 et m'a permis de réaliser les contributions scientifiques que nous rencontrerons en seconde partie.

16. Notez que les enregistrements ont été réalisés dans une pièce isolée à l'intérieur de l'hôpital, ce qui limite le nombre de bruits environnementaux présents, mais reste un bon indicateur d'environnement possible au sein des hôpitaux.

Deuxième partie

Contributions scientifiques pour une approche clinique

Introduction de partie

Dans cette deuxième partie, nous abordons plus particulièrement mes contributions scientifiques pour une approche clinique. Le chapitre 4 cherche à répondre à la question : « quelles sont les modélisations utilisables sachant qu'il y a peu de données pour l'apprentissage ? ». En effet, la taille du corpus cancer ne nous permet pas de procéder à des méthodologies supervisées classiques. Un panorama des méthodes utilisables avec peu de données est dressé, puis une première expérimentation utilisant une approche « few-shots » est proposée sur les voix de personnes présentant des cancers des voies aérodigestives supérieures.

Ensuite, nous détaillons nos expériences utilisant des modèles autosupervisés en chapitre 5. Cela va de l'observation de projections de données, à la création d'une mesure entropique, établie sur un modèle appris en auto-supervision. Nos résultats seront analysés et comparés

Pour finir, dans le chapitre 6, nous présentons le développement de notre application pour tablette permettant de mesurer la sévérité de patients à l'aide de notre meilleure expérimentation. Nous y présentons les enjeux, les fonctionnalités et la réalisation effectuée pour cette application.

4

Approches supervisées

La majorité des systèmes vocaux actuels de pointe utilisent des réseaux neuronaux profonds (DNN) [17]. Nous avons résumé les approches formant l'état de l'art pour la reconnaissance automatique de la parole dans le tableau 4.1 et le tableau 4.2, pour la reconnaissance des émotions dans le tableau 4.3 et pour la reconnaissance du locuteur dans le tableau 4.4. Ces tableaux montrent que le fait de disposer de plus de données n'entraîne pas toujours de meilleurs résultats. Néanmoins, l'utilisation d'une plus grande quantité de parole pour le pré-entraînement d'un modèle non supervisé (tel que le modèle libri-light [Kahn et al., 2020] 60k heures) et de plus grands modèles (modèle ayant plus de paramètres) permet d'obtenir des résultats de pointe sur la parole non altérée. Il est donc utile d'augmenter le nombre de paramètres ou la quantité de données lorsque c'est possible.

Il est donc difficile d'entraîner un système de l'état de l'art sur un corpus de parole avec peu de ressources. Ce qui est notre cas avec le corpus cancer. L'acquisition de nouvelles données et/ou d'expertise est longue et coûteuse. Par conséquent, ceci est exclu pour ma thèse. Dans ce chapitre, nous étudions d'abord les systèmes de reconnaissance automatique de la parole les plus récents. Ensuite, nous essaierons un aperçu des techniques et des tâches nécessitant moins de données. Dans le dernier sous-chapitre, nous étudions les techniques à faible nombre d'occurrences en interprétant la parole insuffisamment documentée comme un problème à faible nombre d'occurrences. Dans ce sens, nous proposons une vue d'ensemble des techniques à faible nombre d'occurrences et la possibilité d'utiliser ces techniques pour notre corpus.

17. En reconnaissance de la parole : https://paperswithcode.com/task/speech-recognition;
En reconnaissance du locuteur : https://paperswithcode.com/task/speaker-verification et https://paperswithcode.com/task/speaker-recognition;
En reconnaissance d'émotion https://paperswithcode.com/task/speech-emotion-recognition

TABLE 4.1 – Résultat état de l'art sur test-clean de librispeech avec la quantité utilisée en pré-entrainement. Toutes les approches ont utilisé le corpus d'entrainement de 960h. Certains des résultats des approches autosupervisées proviennent de [Yang et al., 2021].

Type de modèle	Quantité de données utilisée		WER
	en pré-entrainement	en entrainement	
PASE+ [Ravanelli et al., 2020]	50h	960h	16,6
Wav2Vec2.0 [Baevski et al., 2020]	960h		4,8
	60 000h		3,1
HuBERT [Hsu et al., 2021]	960h		4,8
	60 000h		2,9
Modèle hybride [Lüscher et al., 2019]	-		2,7
Supervisé de bout en bout [Kim et al., 2019]	-		2,4
Wav2Vec2.0 utilisant les conformers et spec augment [Zhang et al., 2020]	60 000h		1,4
Wav2Vec utilisant BERT XXL [Chung et al., 2021]	60 000h		1,4

TABLE 4.2 – Résultat état de l'art sur test-other de librispeech avec la quantité utilisée en pré-entrainement. Toutes les approches ont utilisé le corpus d'entrainement de 960h. Certains des résultats des approches autosupervisées proviennent de [Yang et al., 2021].

Type de modèle	Quantité de données utilisée		WER
	en pré-entrainement	en entrainement	
Supervisé de bout en bout [Kim et al., 2019]	-	960h	8,3
Modèle hybride [Lüscher et al., 2019]	-		5,7
Wav2Vec2.0 utilisant les conformers et spec augment [Zhang et al., 2020]	60 000h		2,6
Wav2Vec utilisant BERT XXL [Chung et al., 2021]	60 000h		2,5

TABLE 4.3 – Résultat état de l'art sur IEMOCAP sur la tâche à quatre émotions (joie, neutre, colère, tristesse) avec la quantité utilisée en pré-entrainement. Toutes les approches ont utilisé le corpus d'entrainement de 12h. Les résultats des approches autosupervisées proviennent de [Yang et al., 2021].

Type de modèle	Quantité de données utilisée		Précision
	en pré-entrainement	en entrainement	
PASE+ [Ravanelli et al., 2020]	50h	12h	57,9
Wav2Vec2.0 [Baevski et al., 2020]	960h		63,4
	60 000h		65,6
HuBERT [Hsu et al., 2021]	960h		64,9
	60 000h		67,6
Approche multi tâches [Li et al., 2019]	-	+ les labels de l'autre tâche	81,6
DAAN [Lian et al., 2020]	1 milliard de mots pour le modèle lexical		82,7

TABLE 4.4 – Résultat état de l'art sur VoxCeleb1 avec la quantité utilisée en pré-entrainement. Toutes les approches ont utilisé le corpus d'entrainement de 350h. Les résultats des approches autosupervisées proviennent de [Yang et al., 2021].

Type de modèle	Quantité de données utilisée		Précision
	en pré-entrainement	en entrainement	
PASE+ [Ravanelli et al., 2020]	50h	350h	38,0
Wav2Vec2.0 [Baevski et al., 2020]	960h		75,2
	60 000h		86,1
HuBERT [Hsu et al., 2021]	960h		81,4
	60 000h		90,3
AutoSpeech [Ding et al., 2020]	-		87,7

4.1 Supervision classique

4.1.1 État des lieux

Les performances des systèmes de traitement automatique de la parole se sont considérablement améliorées lors des dernières années, en particulier les systèmes de reconnaissance automatique de la parole. C'est également le cas pour d'autres tâches de traitement de la parole, comme l'identification du locuteur ou la classification des émotions. Ce succès a été rendu possible par la grande quantité de données annotées disponibles, combinée à l'utilisation intensive de techniques d'apprentissage profond et à la capacité des unités de traitement graphique modernes. Certaines modélisations sont déjà déployées pour un usage quotidien, comme les assistants personnels des smartphones, les enceintes connectées, etc. Néanmoins, des défis restent à relever pour les systèmes de traitement automatique de la parole.

Ils manquent de robustesse pour traiter des vocabulaires étendus dans un environnement réel : cela inclut les bruits environnementaux, la distance par rapport au locuteur, les réverbérations, manque de robustesse aux variations de la parole et autres altérations [Sahu et al., 2018]. Certains défis, comme CHiME [Barker et al., 2018], fournissent des données pour permettre à la communauté d'essayer de traiter certains de ces problèmes. On cherche des moyens d'améliorer la généralisation des modèles modernes en évitant d'inclure d'autres données annotées pour chaque environnement possible.

Les techniques de l'état de l'art (SOTA) pour la plupart des tâches vocales nécessitent de grands ensembles de données. En effet, avec les systèmes modernes de traitement de la parole DNN, disposer de plus de données implique généralement de meilleures performances. Le TED-LIUM 3 (de [Hernandez et al., 2018], avec 452 heures) fournit plus du double des données du jeu de données TED-LIUM 2. Les auteurs obtiennent donc de meilleurs résultats en entraînant leur modèle sur TED-LIUM 3 qu'en entraînant leur modèle à l'aide des données TED-LIUM 2. Cette amélioration des performances des systèmes ASR est également observée avec le jeu de données LibriSpeech, de [Panayotov et al., 2015]. V. Panayotov et al. ont obtenu de meilleurs résultats sur l'ensemble de test du Wall Street Journal (WSJ) en entraînant un modèle à l'aide de l'ensemble de données LibriSpeech (1 000 heures) qu'avec l'ensemble d'entraînement du WSJ (82 heures) [Panayotov et al., 2015].

Ce phénomène, selon lequel disposer de plus de données permet d'obtenir de meilleures performances, est également observable en identification du locuteur avec le jeu de données VoxCeleb 2 par rapport au jeu de données VoxCeleb [Chung et al., 2018] : les auteurs ont augmenté le nombre de phrases de 100 000 à un million et augmenté le nombre d'individus de 1251 à 6112 en comparaison à la version précédente de VoxCeleb. Ils ont ainsi obtenu de meilleures performances qu'en entraînant leur modèle avec le précédent jeu de données VoxCeleb.

Avec les langues sous-ressourcées (telles que [Deka et al., 2018]) et/ou certaines tâches (détection de pathologies à partir de signaux vocaux), nous manquons de grands ensembles de données [Latif et al., 2020]. Par sous-ressourcé, nous entendons des ressources numériques limitées (corpus acoustiques et textuels limités) et/ou un manque d'expertise linguistique. Pour une définition plus précise et les détails du problème, voir [Besacier et al., 2014]. Certaines tâches vocales non conventionnelles telles que la détection de maladies (telles que la maladie de Parkinson, la gravité du cancer ORL et autres) à l'aide de l'audio sont des exemples de tâches manquant de ressources [Latif et al., 2020]. La formation de modèles de réseaux neuronaux profonds dans de tels contextes est un défi pour ces ensembles de données vocales sous-ressourcé. C'est particulièrement le cas pour les tâches impliquant un grand vocabulaire. M. Moore et al. ont montré que les systèmes ASR récents sont mal adaptés à la parole altérée [Moore et al., 2018] et M. B. Mustafa et al. ont montré les difficultés à adapter de tels modèles avec des quantités limitées de données [Mustafa et al., 2014]. Entraîner un système ASR sur une nouvelle langue, adapter un système ASR sur une parole pathologique ou réaliser une identification du locuteur (avec une voix altérée) avec peu d'exemples restent des tâches compliquées [Moore et al., 2018, Latif et al., 2020].

Ce sous-chapitre se concentrera sur la façon de former des modèles de réseaux neuronaux profonds (DNN) avec peu de ressources pour des données vocales avec des signaux mono non chevauchants. Par conséquent, nous examinerons d'abord les techniques d'ASR SOTA qui utilisent une grande quantité de données (sous-chapitre 4.1.2). Ensuite, nous examinerons les techniques et les tâches vocales (identification du locuteur, reconnaissance des émotions) nécessitant moins de données que les techniques SOTA (sous-chapitre 4.1.3). Nous examinerons également le traitement de la parole pathologique pour l'ASR avec des techniques d'adaptation (sous-chapitre 4.1.3).

4.1.2 Système de traitement automatique de la parole

Dans ce sous-chapitre, nous allons passer en revue les systèmes d'ASR, de reconnaissance d'émotion et d'identification du locuteur obtenant l'état de l'art et utilisant des modèles multiples et des modèles de bout en bout. Ici, nous nous concentrons sur les séquences monophoniques $\mathbf{x} = [x_1, x_2, \ldots, x_n]$ où x_i peuvent être des paramètres acoustiques ou des échantillons audio. Les systèmes d'ASR consistent à faire correspondre $\mathbf{x}$ à une séquence de mots $\mathbf{y} = [y_1, y_2, \ldots, y_u]$ (où $u \leq n$). Alors que les systèmes de reconnaissance des locuteurs et des émotions transforment la séquence $\mathbf{x}$ en un résultat unique y représentant une classe. Les systèmes examinés pour mon travail ont été évalués grâce au taux d'erreurs sur les mots (WER) comme mesure pour les systèmes d'ASR et la précision pour les systèmes de reconnaissance d'émotion et pour les systèmes d'identification du locuteur.

Multi-modèles

Une approche multi-modèles consiste à résoudre un problème à l'aide de plusieurs modèles. Ces modèles sont conçus pour résoudre des sous-tâches (liées au problème) et la tâche ciblée. La configuration minimale est avec deux modèles (disons f et g) pour résoudre une tâche donnée. Classiquement, pour la tâche ASR, nous pouvons d'abord entraîner un modèle acoustique (un classificateur de phonèmes ou une unité sonore équivalente), puis entraîner un modèle de langage qui produit la séquence de mots souhaitée. Par conséquent, nous avons :

$$\hat{\mathbf{y}} = f(g(\mathbf{x})) \tag{4.1}$$

avec f étant le modèle de langage prédisant la séquence de sortie $\hat{\mathbf{y}}$ (qui peut être différente de la séquence réelle $\mathbf{y}$) et g étant le modèle acoustique. Notez que pour la reconnaissance des émotions et du locuteur, la sortie de f est $\hat{y}$ au lieu de $\hat{\mathbf{y}}$ et n'est pas nécessairement un modèle de langage. Les deux peuvent être entraînés séparément ou conjointement. En général, les modèles hybrides sont utilisés comme modèles acoustiques.

Les modèles hybrides consistent à utiliser des modèles probabilistes avec des modèles neuronaux. Les modèles de mélange de lois gaussiennes (GMM) sont un exemple de modèles probabilistes. Un modèle hybride populaire et efficace est le DNN-Hidden Markov Model (DNN-HMM). Ce type d'architecture est généralement utilisé pour créer un modèle acoustique. Ce modèle acoustique est combiné avec un modèle de langage (LM) qui transforme les phonèmes en une séquence de mots. Lüscher et al. ont utilisé des DNN-HMM combinés à un LM pour obtenir SOTA sur l'ensemble de tests LibriSpeech (ensemble de tests officiels augmentés) [Lüscher et al., 2019]. Ce modèle traite les coefficients cepstraux de fréquence Mel (MFCC), calculés sur les signaux audio. Leur meilleur LM consistait à utiliser le Transformer de [Vaswani et al., 2017]. Les Transformers sont des modèles autorégressifs (dépendant des sorties précédentes du modèle) utilisant des mécanismes d'attention dite douce. L'attention douce consiste à déterminer un aperçu g, qui est une sélection de caractéristiques de l'entrée x permettant de filtrer les informations non utiles. Ainsi, parmi tous les aperçus possibles, g est calculé comme suit :

$$g = \sum_{g' \in x} g' Pr(g'|a) \tag{4.2}$$

avec x étant les données d'entrée et a les paramètres d'attention. Leur meilleur modèle hybride a obtenu un WER de 5,7% pour l'ensemble test-other (test avec des enregistrements dans des environnements bruités) et de 2,7% pour l'ensemble test-clean (test avec des enregistrements dans des environnements non bruités).

L'approche hybride est également utilisée dans la reconnaissance des émotions et obtient l'état de l'art sur IEMOCAP grâce au réseau neuronal adversarial de domaine dépendant du contexte (DAAN) [Lian et al., 2020]. La base de données IEMOCAP [Busso et al., 2008] a été modifiée pour obtenir un problème d'émotion à quatre classes. Ces émotions sont la colère, la joie, la neutralité et la tristesse. Le réseau DAAN consiste à utiliser en entrée un modèle lexical (pré-entraîné sur un milliard de mots) et des caractéristiques audio (telles que MFCC et énergie) qui représentent 6373 caractéristiques pour chaque image. Ils fusionnent ces entrées grâce à l'attention. Ensuite, des GRU multicouches sont utilisés. Notez que leur approche utilise deux tâches à apprendre (reconnaissance des émotions et reconnaissance du domaine). Une

telle approche nécessite de multiples étiquettes, que la base de données IEMOCAP fournit. En procédant ainsi, ils ont obtenu une précision de 82,7% pour toutes les émotions.

Systèmes de bout en bout

Dans les approches de bout en bout, l'objectif est de déterminer un modèle f qui peut effectuer la mise en correspondance :

$$\hat{\mathbf{y}} = f(\mathbf{x}) \tag{4.3}$$

L'apprentissage se fera directement de la séquence $\mathbf{x}$ à la séquence $\mathbf{y}$ désirée. Notez que pour la reconnaissance des émotions et du locuteur, la sortie de f est $\hat{y}$ au lieu de $\hat{\mathbf{y}}$. Seules les méthodes supervisées peuvent fonctionner de bout en bout pour résoudre les tâches vocales qui nous intéressent.

Dans les systèmes ASR, l'état de l'art sur l'ensemble test-clean de LibriSpeech est obtenu par [Kim et al., 2019]. En comparaison à [Lüscher et al., 2019], ils ont utilisé la perturbation de la longueur du tract vocal comme entrée de leur modèle de bout en bout. Le modèle est établi sur l'architecture encodeur-décodeur utilisant une mémoire à court et long terme (LSTM) empilée pour l'encodeur et une LSTM combinée à l'attention douce pour le décodeur [Kim et al., 2019]. Ils ont obtenu un WER de 2,44% sur l'ensemble test-clean et de 8,29% sur l'ensemble test-other. Ces résultats sont proches de ceux de [Lüscher et al., 2019] (meilleurs résultats du modèle hybride) et montrent que les approches de bout en bout sont compétitives face aux approches multi-modèles.

En reconnaissance d'émotions, Li et al. ont obtenu l'état de l'art avec une version modifiée de la base de données IEMOCAP [Busso et al., 2008] résultant en un problème à quatre classes. Ces émotions sont la colère, la joie, la neutralité et la tristesse. Ils ont utilisé un système multi-tâches de bout en bout avec uniquement des tâches supervisées : l'identification du genre et l'identification des émotions [Li et al., 2019]. Le modèle résultant a atteint une précision globale pour la tâche d'émotion (qui est la cible principale) de 81,6% et une précision moyenne de chaque catégorie d'émotion de 82,8%. L'utilisation d'une telle approche leur a permis d'obtenir des résultats équilibrés à partir de données non équilibrées. Néanmoins, l'utilisation de tâches supervisées uniquement nécessite plusieurs vérités fondamentales pour l'ensemble de données ciblé.

Dans l'identification du locuteur, l'architecture autoSpeech [Ding et al., 2020] obtient l'état de l'art sur le jeu de données VoxCeleb1 [Nagrani et al., 2017]. Leur approche consiste en un algorithme qui recherche automatiquement la meilleure architecture de réseau de neurones convolutifs pour résoudre la tâche. Avec leur approche, ils ont obtenu une précision de 87,66%.

4.1.3 Techniques nécessitant moins de données

Certaines techniques nécessitent moins de données que les techniques vues dans le précédent sous-chapitre. Dans ce sous-chapitre, nous énumérons les principales façons de surmonter (à notre connaissance) l'absence de grands ensembles de données constitués de parole pathologique. Nous n'aborderons pas les techniques semi-supervisées qui utilisent une grande quantité de données non supervisées.

Augmentation des données

La première façon d'exploiter le manque de données est d'augmenter artificiellement le nombre de données. Pour ce faire, l'approche classique consiste à ajouter du bruit ou de la déformation, comme dans [Park et al., 2019]. Les auteurs obtiennent des résultats proches de l'état de l'art sur Librispeech (1 000 heures de [Panayotov et al., 2015]) avec un modèle de bout en bout. De plus, ils obtiennent l'état de l'art sur SwitchBoard (300 heures de [Godfrey et al., 1992]) avec un WER de 6,8% sur Switchboard et 14,1% sur la partie CallHome en utilisant la fusion peu profonde et leur augmentation de données. Néanmoins, il s'agit d'augmentations d'altérations du signal et certaines d'entre elles nécessitent des données audio supplémentaires (comme l'ajout de bruit).

D'autres approches utilisent des modèles génératifs pour obtenir de nouveaux échantillons, comme dans [Chatziagapi et al., 2019,Jiao et al., 2018]. Parmi ces approches, l'utilisation de réseaux adversaires génératifs (GAN) conditionnels permet de générer de nouveaux échantillons tout en contrôlant les labels des exemples générés [Mirza and Osindero, 2014]. En procédant ainsi, il est aisé d'équilibrer un jeu de données initial et d'obtenir de meilleurs résultats. Il est à noter que les GANs à convolution profonde ont également été utilisés pour générer de la parole dysarthrique et améliorer des résultats existant sans une telle augmentation [Jiao et al., 2018].

Transposition de domaine

Une autre façon de tirer parti du manque de données est d'utiliser la transposition de domaine de données. L'idée consiste à transposer les caractéristiques de la parole (ou des paramètres acoustiques) d'un domaine (tel que des spectrogrammes contenant la parole + des bruits) à un autre domaine pour réduire la complexité des données (tel que des spectrogrammes contenant uniquement la parole). En apprentissage automatique, la complexité d'une donnée est définie par la taille du modèle nécessaire pour répliquer les données, la taille de l'encodage le plus court possible (qui permet de reconstruire les données initiales) et le taux d'erreur du meilleur modèle possible compte tenu d'une tâche [Li and Abu-Mostafa, 2006].

Voici quelques exemples récents sur le traitement de la parole :

- L'utilisation de GAN pour déréverbérer des signaux vocaux [Wang et al., 2018]. Dans leur travail, le générateur est utilisé comme une fonction de mappage pour convertir les signaux réverbérés en signaux de parole déréverbérés.
- L'utilisation de GAN pour rendre des signaux contenant des troubles de la parole en des signaux de paroles non altérés par des troubles de la parole [Chen et al., 2019]. Les auteurs ont effectué une conversion vocale grâce à un GAN avec un contrôleur mappant la parole altérée vers un espace de représentation z. z étant l'entrée du générateur qui est utilisé comme une fonction de mappage pour avoir des signaux de parole non altérés.
- L'utilisation de Cycle GAN (cadre conçu pour le transfert de domaine) comme améliorateur audio [Zhao et al., 2019]. Leur résultat constitue l'état de l'art sur le jeu de données CHiME-4 (jeux de données d'audio très bruité).

Modèles nécessitant moins de paramètres

Le fait de disposer de moins de données signifie qu'un surapprentissage peut se produire si les modèles de réseaux neuronaux nécessitent trop de paramètres. C'est pourquoi certaines techniques ont essayé des modèles nécessitant moins de paramètres. Nous présentons ici quelques techniques récentes qui nous semblent intéressantes :

- L'utilisation de couches SincNet, de [Ravanelli and Bengio, 2018], pour remplacer les convolutions 1D classiques sur de l'audio brut. Ici, au lieu de nécessiter $window\_size$ paramètres (avec $window\_size$ étant la taille de la fenêtre de la convolution 1D) par filtre, nous n'avons besoin que de deux paramètres par filtre pour chaque taille de fenêtre. Ces deux paramètres représentent indirectement les valeurs de la bande passante à haute et basse énergie.
- L'utilisation de LightGRU (LiGRU), de [Ravanelli et al., 2018], fondé sur les Gated Recurrent Unit (GRU). Les LiGRU sont une simplification du cadre GRU compte tenu de certaines hypothèses concernant le signal vocal. Ils ont supprimé la porte de réinitialisation des GRU et utilisé la fonction d'activation $ReLU$ (combinée à la normalisation par batch [Ioffe and Szegedy, 2015]) au lieu de la fonction d'activation $\tanh$.
- L'utilisation des réseaux neuronaux quaternion, de [Parcollet et al., 2018], pour le traitement de la parole. La formulation en quaternions permet de fusionner quatre dimensions en une seule, ce qui entraîne une réduction drastique des paramètres requis dans leurs expériences (presque quatre fois moins).

Approche multi-tâches

Les modèles multi-tâches peuvent être considérés comme une extension de l'architecture Encodeur-Décodeur où l'on retrouve un décodeur par tâche avec un encodeur partagé (comme dans la Figure 4.1) pour tous les décodeurs. Ces tâches sont ensuite formées conjointement avec des algorithmes classiques de type feed-forward. L'objectif de l'apprentissage multi-tâches est d'obtenir un encodeur qui produit suffisamment d'informations pour résoudre chaque tâche. Il peut donc potentiellement améliorer les performances de chaque tâche face aux architectures mono-tâches. C'est également un moyen d'obtenir un encodeur plus représentatif pour une même quantité de données.

Pascual et al. ont utilisé une combinaison de tâches auto-supervisées pour résoudre ce problème et ont utilisé l'encodeur résultant pour l'apprentissage par transfert [Pascual et al., 2019]. Ils ont récemment amélioré ce travail dans [Ravanelli et al., 2020] où ils utilisent plus de tâches, une unité récurrente au-dessus de l'encodeur et des mécanismes de débruitage utilisant des techniques d'augmentation de données multiples sur leur système.

Apprentissage par transfert

Les techniques d'apprentissage par transfert consistent à utiliser un modèle pré-entraîné et en l'utilisant comme un extracteur de caractéristiques ou avec ses paramètres comme des initialisateurs (qui peuvent être affinés) pour résoudre un problème/tâche connexe.

Speech2Vec [Chung and Glass, 2018] est une adaptation de la technique de représentation Word2vec pour le traitement du langage naturel (NLP).

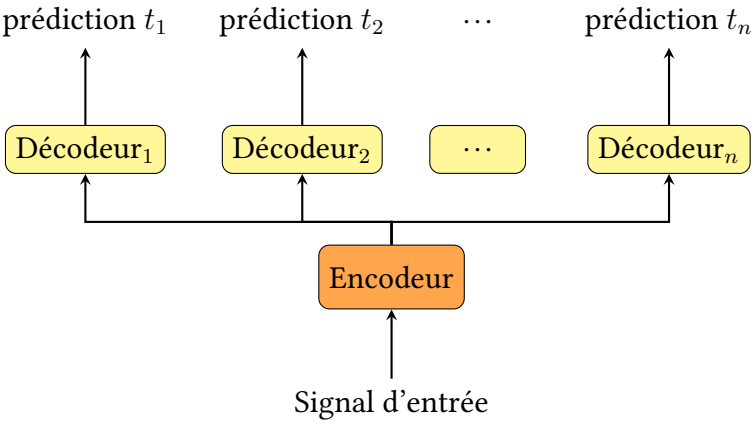


FIGURE 4.1 – Illustration de l'architecture multi-tâches. La sortie de l'encodeur est donnée à chaque décodeur pour produire la prédiction pour chaque tâche t_i.

Contrastive Predictive Coding (CPC de [Oord et al., 2018]) est une architecture pour l'apprentissage de la représentation audio non supervisée utilisant une architecture à 2 niveaux combinée à une fonction de coût auto-supervisée. Les auteurs ont obtenu de meilleurs résultats en transférant les paramètres obtenus sur une tâche d'identification de locuteur et une tâche de classification de phonèmes (sur le jeu de données LibriSpeech) qu'avec l'utilisation de paramètres MFCC.

Certaines tâches binaires du modèle multi-tâches de Pascual et al. utilisent le codage prédictif comme dans les CPC. Ces derniers ont développé un modèle multi-tâches non auto-supervisées pour obtenir de meilleurs encodeurs pour l'apprentissage par transfert [Pascual et al., 2019]. Ils l'ont appliqué à de multiples tâches et ont obtenu des résultats acceptables sur l'identification du locuteur (avec VTCK [Yamagishi et al., 2019]), la reconnaissance des émotions (avec INTERFACE [Hozjan et al., 2002]) et l'ASR (avec TIMIT [Garofolo et al., 1993]).

De nos jours, les approches CPC telles que Wav2Vec [Schneider et al., 2019] (avec sa dernière version Wav2Vec2 [Baevski et al., 2020]) et HuBERT [Hsu et al., 2021] représentent l'état de l'art des techniques auto-supervisées. Wav2Vec et Wav2Vec2 consistent à apprendre des représentations à partir de la forme d'onde de l'audio avec le cadre CPC. Wav2Vec2 combine des couches convolutionnelles pour traiter la forme d'onde et obtenir une représentation de faible niveau, qui est transmise aux couches d'un transformer pour obtenir une représentation contextuelle. Alors que HuBERT utilise un encodeur BERT [Devlin et al., 2019] (qui est un transformateur) et sa fonction de perte dans un cadre CPC.

Ces approches ont été comparées dans des benchmarks [Yang et al., 2021] (entre autres). Pour l'identification du locuteur, le meilleur modèle a obtenu une précision de 90,33% sur le jeu de test VoxCeleb [Nagrani et al., 2020]. Ensuite, pour la reconnaissance des émotions, le meilleur modèle a obtenu une précision de 67,62% sur le jeu de tests IEMOCAP [Busso et al., 2008]. Enfin, pour l'ASR, le meilleur modèle a obtenu un taux d'erreur sur les mots de 2,94% sur le jeu d'essai « test-clean » de librispeech [Panayotov et al., 2015]. Une vue d'ensemble plus détaillée (avec un benchmark complet sur plusieurs tâches vocales) des approches auto-supervisées est disponible dans [Yang et al., 2021]. Notez que leur benchmark n'inclut pas les données de parole pathologique ou les tâches liée à de la parole pathologique.

Les avantages des réseaux pré-entraînés pour l'apprentissage par transfert diminuent lorsque la tâche cible diverge de la tâche originale du réseau pré-entraîné [Yosinski et al., 2014]. Pour remédier à cela, Van den Oord et al. ont tenté d'avoir des tâches génériques avec leur approche non supervisée, et ils ont obtenu des résultats prometteurs [Oord et al., 2018]. Les avantages de l'apprentissage par transfert diminuent également lorsque la dissemblance entre les ensembles de données augmente [Yosinski et al., 2014]. Ce problème peut décourager l'utilisation de l'apprentissage par transfert pour certains discours pathologiques. Cependant, la parole dysarthrique et la parole accentuée semblent semblables à la parole du jeu de données LibriSpeech, selon [Shor et al., 2019]. Shor et al. ont utilisé avec succès l'apprentissage par transfert pour améliorer leurs résultats avec un jeu de données de 36,7 heures.

Néanmoins, Mustapha et al. ont montré que les paramètres acoustiques de la parole non altérée et altérée sont très différentes [Mustafa et al., 2014]. Lorsque peu de données sont disponibles, ces problèmes peuvent être critiques.

4.1.4 Nos expériences sur notre corpus cancer

Suite à la revue des méthodes vues en début de chapitre, nous nous sommes posé la question de l'applicabilité de ces techniques pour notre corpus. En effet, pour notre corpus de données cancer, nous ne pouvons pas utiliser certaines techniques. Le transfert de domaine des données et l'augmentation de données risquent de modifier le signal initial des participants. Ceci peut potentiellement invalider notre vérité terrain (car les experts risquent de donner des scores différents si la parole des participants est modifiée). Pour garantir que ces altérations n'invalident pas nos annotations, il serait nécessaire de réaliser des campagnes d'annotations (coûteuses en temps et argent). Il serait plus avisé de réaliser plus d'enregistrements de données patients et contrôles au vu des quantités de données de notre corpus.

Enfin, nous ne disposons pas de données non labellisées dans notre corpus initial ou provenant d'autres corpus français. Ainsi, nous ne pouvons pas dans notre cas utiliser des approches semi-supervisées. Par conséquent, nous nous sommes tournés vers l'apprentissage utilisant moins de paramètres et l'apprentissage par transfert.

4.1.5 Apprentissage avec modèle contenant peu de paramètres

Pour l'apprentissage avec peu de paramètres, nous nous sommes tournés sur les tâches de lecture et de description. Nous avons tenté de construire un modèle fondé sur une régression directe entre le score perceptif de sévérité et des segments d'une seconde de signal de parole. Au préalable, nous avons utilisé notre détecteur d'activité vocale (vu dans le sous-chapitre 5.2.2).

Nous avons essayé un modèle établi sur les MFCC et un autre établi sur les sincnet. Dans les deux cas, nous avons utilisé des couches de décision identiques. Cette architecture est résumée en table 4.5 en utilisait un optimiseur adam et avec une fonction de coût fondée sur l'erreur quadratique moyenne (MSE) avec le score donné par le jury d'experts. Pour les MFCC et pour les sincnet, nous avons utilisé 13 filtres avec des fenêtres de 25 ms et un recouvrement avec des pas de 10 ms. L'idée de ces choix est d'augmenter artificiellement le nombre d'entrées de notre modèle tout en utilisant peu de paramètres.

Cependant, nous avons obtenu des performances médiocres (avec une MSE d'environ 7 pour un score allant de 1 à 10). Après analyse, nous avons constaté que le modèle avait surentrainé (malgré l'utilisation d'early stopping). Par conséquent, cette première approche n'est pas concluante. Elle peut être liée aux entrées choisies, à la formalisation de la tâche ou au manque de données. Pour la suite de ces expériences, nous avons d'abord supposé que l'entrée n'était pas adaptée pour la tâche, c'est pourquoi nous avons choisi de faire de l'apprentissage par transfert. sur les tâches de lecture et de description. Nous avons tenté de construire un modèle fondé sur une régression directe entre le score perceptif de sévérité et des segments d'une seconde de signal de parole. Au préalable, nous avons utilisé notre détecteur d'activité vocale vu dans le sous-chapitre 5.2.2.

TABLE 4.5 – Architecture du réseau avec peu de paramètres utilisés pour nos premières expériences.

N° couche	Type de couche	Paramètres
0	donnée en entrée	MFCC ou sincnet ou PASE+
1	Normalisation des dimensions en entrée	Rélisé avec une layer norm sur les données d'entrainement
2	linéaire	128 filtres
3	activation	ReLU
4	linéaire	64 filtres
5	activation	ReLU
6	linéaire	32 filtres
7	activation	ReLU
8	linéaire	1 filtre représentant le score

4.1.6 Apprentissage par transfert

Pour l'apprentissage par transfert, nous avons utilisé l'encodeur du modèle PASE+ en entrée de l'architecture vu en table 4.5. Pour ce transfert, nous avons figé les paramètres de l'architecture de PASE+. Néanmoins, nous avons obtenu des résultats similaires que ceux précédemment effectués.

Suite à ce constat, nous pensons que le problème vient d'un manque de données. En effet, dans la base de données cancer, les valeurs des scores de sévérité ne couvrent pas uniformément toutes les plages de valeurs : il y a moins de représentants à scores faibles que de représentants de scores élevés. Néanmoins, l'acquisition de résultats pour toutes les plages est une tâche difficile, voire impossible. De plus, une autre raison possible de l'échec de l'apprentissage par transfert est que nous utilisons une tâche divergente de la tâche ayant permis d'apprendre le modèle que l'on transfère. En effet, ces conditions selon [Yosinski et al., 2014] conduisent souvent à des résultats non satisfaisants.

4.1.7 Résumé et perspectives

La nécessité de créer des ensembles d'entraînement et de tests pour les tâches de régression implique que peu de données sont disponibles pour l'entrainement. Ne disposant pas de suffisamment de données, cela est possiblement le principal problème pour l'apprentissage par transfert. Ainsi, les techniques « traditionnelles » pour pallier le manque de données n'ont pas fonctionné pour notre cas. C'est pourquoi pour la suite de mes travaux, nous nous sommes tournés vers des techniques d'images fonctionnant sur des corpus avec très peu de données : les techniques de "few-shot".

4.2 Supervision avec du few-shot

L'apprentissage Few-shot consiste à entraîner un modèle en utilisant k-shot (où shot signifie un exemple par classe), où $k \geq 1$ et k est un nombre faible. L'entraînement d'un système ASR sur une nouvelle langue, l'adaptation d'un système ASR sur une parole pathologique ou l'identification d'un locuteur avec peu d'exemples restent des tâches compliquées. Nous pensons que les techniques de "few-shot" peuvent être utiles pour aborder ces problèmes de manque de ressources.

4.2.1 Few-shot état des lieux

Nous avons passé en revue les techniques nécessitant une grande quantité de données. Suite à nos premières expériences, nous nous sommes rendu compte que nous ne disposons pas de suffisamment de données pour utiliser de telles techniques. Cependant, celle-ci n'est pas toujours disponible, comme pour la parole pathologique [Latif et al., 2020]. Il est à noter que de grosses entreprises comme Google tentent d'acquérir davantage de données de cette nature [18]. Malgré tout, l'acquisition de telles données peut être assez coûteuse et prend beaucoup de temps. Mustafa et al. recommandent l'utilisation de techniques adaptatives pour résoudre le problème de la quantité limitée de données dans un tel cas [Mustafa et al., 2014].

Nous pensons que les techniques few-shot peuvent être une autre solution à ce problème. Néanmoins, des tâches comme la reconnaissance de parole pathologique ou l'identification de certains dialectes (dont peu d'exemples sont disponibles) restent difficiles à entraîner avec les techniques obtenant les meilleures performances. En effet, ces techniques ont été créées pour être apprises sur de grands ensembles de données vocales. C'est pourquoi nous étudions les techniques suivantes, utilisant un petit nombre d'exemples et envisageons les adaptations nécessaires pour les utiliser sur des jeux de données et tâches de parole.

Notations sur les quelques points

Considérons une distribution P depuis laquelle nous tirons des épisodes indépendants identiquement distribués (iid) ($\mathcal{E}$ ou ensembles de données). L'ensemble $\mathcal{E}$ est composé d'un

18. https://blog.google/outreach-initiatives/accessibility/impaired-speech-recognition/

ensemble de support $\mathcal{S}$, de données non étiquetées $\bar{\mathbf{x}}$ et d'un ensemble de requêtes $\mathcal{Q}$. L'ensemble de support correspond aux échantillons supervisés auxquels le modèle à accès :

$$\mathcal{S} = \{(x_1, y_1), \dots (x_s, y_s)\} \tag{4.4}$$

avec x_i les échantillons et y_i les étiquettes correspondantes, telles que $y_i \in \{1, 2, \dots, K\}$ et K étant le nombre de classes apparaissant dans P.

L'ensemble de requêtes est composé d'échantillons permettant de classer $\hat{\mathbf{x}}$ avec $\hat{\mathbf{y}}$ étant la vérité terrain correspondante.

Pour résumer, les épisodes tirés de P ont la forme suivante :

$$\begin{aligned} \mathcal{E} = \{&\mathcal{S} = \{(x_1, y_1), \dots, (x_s, y_s)\}, \\ &\bar{\mathbf{x}} = (\bar{x}_1, \dots, \bar{x}_r), \\ &\mathcal{Q} = \{(\hat{x}_1, \hat{y}_1), \dots, (\hat{x}_t, \hat{y}_t)\} \end{aligned} \tag{4.5}$$

avec s, r et t des valeurs fixes qui représentent respectivement le nombre d'échantillons supervisés pour l'ensemble de support, le nombre d'échantillons non supervisés et le nombre d'échantillons supervisés pour l'ensemble de requêtes.

Dans cette étude, nous nous concentrerons sur les techniques d'apprentissage few-shot, où $r = 0$ (nous ne possédons pas d'exemples non supervisés), $t \geq 1$ et $s = kn$, avec n étant le nombre de fois que chaque étiquette apparaît pour l'ensemble de support et k le nombre de classes sélectionnées parmi P, tel que $k \leq K$. Par conséquent, nous avons un n-shot avec k-ways (ou classes) pour chaque épisode. L'apprentissage à 1-shot n'est qu'un cas particulier de l'apprentissage à quelques shots où $n = 1$. Dans certains cadres d'apprentissage few-shot, nous n'échantillonnons qu'un seul épisode de P et il représente notre tâche.

Dans les sous-chapitres qui suivent, nous passerons en revue les techniques few-shot qui ont eu un impact dans le domaine du traitement de l'image. Nous recenserons les techniques dont la formulation semble convenir au traitement de la parole ainsi que les techniques ayant déjà été utilisées avec succès par la communauté de la parole.

Les réseaux siamois

Les réseaux neuronaux siamois sont conçus pour être utilisés par épisode [Koch et al., 2015]. Ils consistent à mesurer la distance entre deux échantillons et à juger s'ils sont similaires ou non. Par conséquent, un réseau siamois utilise les échantillons de l'ensemble de support $\mathcal{S}$ comme références pour chaque classe. Il est ensuite entraîné à l'aide de toutes les combinaisons d'échantillons de $\mathcal{S} \bigcup \mathcal{Q}$, ce qui permet un entraînement beaucoup plus important que si l'on disposait uniquement de $s+t$ échantillons dans les cadres feedforward classiques. Les réseaux siamois prennent deux échantillons (x_1 et x_2) en entrée et calculent une distance entre eux, comme suit :

$$\phi(x_1, x_2) = \sigma(\sum \boldsymbol{\alpha} |Enc(x_1) - Enc(x_2)|) \tag{4.6}$$

avec Enc étant un DNN encodeur qui représente le signal d'entrée, σ étant la fonction sigmoïde, $\boldsymbol{\alpha}$ des paramètres apprenables qui pondèrent l'importance de chaque composant de l'encodeur et x_1 et x_2 échantillonnés soit depuis l'ensemble de support, soit depuis l'ensemble des requêtes.

Pour définir la classe d'un nouvel échantillon de $\mathcal{Q}$ ou de toute nouvelle donnée, nous devons calculer la distance entre chaque référence de $\mathcal{S}$ et le nouvel échantillon. Un exemple de comparaison entre une référence et un nouvel exemple est illustré dans la Figure 4.2. La classe de la référence présentant la distance la plus faible devient alors la prédiction du modèle.

Pour entraîner un tel modèle, [Koch et al., 2015] ont utilisé cette fonction de perte :

$$\mathcal{L} = \mathbb{E}_{y(x_i)=y(\tilde{x}_j)} \log(\phi(x_i, \tilde{x}_j)) + \mathbb{E}_{y(x_i)\neq y(\tilde{x}_j)} \log(1 - \phi(x_i, \tilde{x}_j))$$

avec $\tilde{\mathbf{x}} = [x_1, \ldots, x_s, \hat{x}_1, \ldots, \hat{x}_t]$ de $\mathcal{S}$ et $\mathcal{Q}$. $y(x)$ est une fonction qui renvoie l'étiquette correspondant à l'exemple x. La dernière couche de ϕ doit être une fonction softmax.

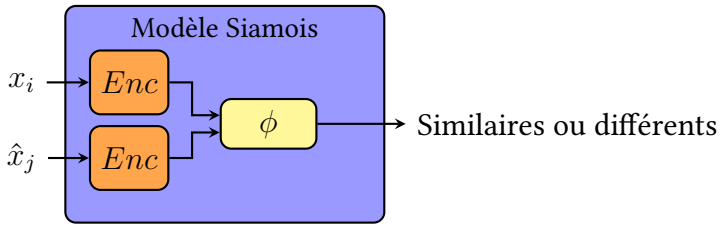


FIGURE 4.2 – Exemple de comparaison entre une référence (x_i) et un nouvel exemple ($\hat{x}_j$) de l'ensemble d'interrogation, où Enc est le même réseau appliqué à la fois à x_i et à $\hat{x}_j$. Le modèle fournit en sortie la distance entre les classes x_i et $\hat{x}_j$.

Eloff et al. ont utilisé une version modifiée de cette technique pour l'apprentissage multimodal, les modalités étant les signaux de parole et d'image [Eloff et al., 2019]. Les signaux vocaux utilisés consistent en des nombres à 11 chiffres (de zéro à neuf et le "oh" spécifique aux anglophones qui l'utilise dans les numéros de téléphone) avec les 10 images correspondantes (oh et zéro donnent les mêmes images). Le problème est d'associer les signaux vocaux aux images correspondantes. Dans leur expérience, le modèle montre une certaine invariance face aux locuteurs (précision de 70,12% $\pm$ 0,68) avec une configuration à 1-shot uniquement, ce qui est un résultat prometteur. Toutefois, à notre connaissance, il n'existe pas encore d'étude concernant uniquement le traitement de la parole.

Toutefois, les réseaux de neurones siamois ne sont pas très adaptés lorsque le nombre de classes K ou que le nombre de shots q devient trop élevé. Cela augmente le nombre de références à comparer et le temps de calcul pour transmettre le modèle. Le problème concerne principalement l'entraînement du modèle. Une fois le modèle entraîné, nous pouvons réduire cet effet en pré-calculant tous les encodages des exemples de l'ensemble de support. Néanmoins, pour la phase d'entraînement cette astuce est impossible et ce problème augmente considérablement le nombre de combinaisons. Ceci peut être vu comme un point positif si l'on n'est pas limité en taille de calculs et que l'on cherche à avoir le jeu de données le plus conséquent possible pour la phase d'entrainement. Malgré cela, cette approche ne semble pas appropriée pour l'ASR de bout en bout avec de grands vocabulaires, comme en anglais (environ 470 000 mots), bien qu'il puisse être suffisant pour des langues comme l'espéranto (environ 16 780 mots). L'autre façon d'utiliser un tel cadre dans les systèmes ASR est de l'utiliser dans des modèles hybrides comme un modèle acoustique, où nous pouvons l'entraîner sur chaque phonème (par exemple 44 phonèmes/sons en anglais) ou des unités sonores plus raffinées.

L'approche des réseaux siamois semble intéressante pour des tâches telles que l'identification du locuteur. En effet, on peut ajouter un nouveau locuteur sans réentraîner le modèle (en supposant que le modèle se soit généralisé) ni changer l'architecture du modèle. Il suffit d'ajouter au moins un échantillon du nouveau locuteur aux références. De plus, la formulation siamoise semble bien adaptée à la vérification du locuteur. Cela nécessite de remplacer la paire $(\mathbf{x}, speaker\_id)$ par la paire $(\mathbf{x}, \mathcal{S}_{top5})$, où $\mathcal{S}_{top5}$ est un ensemble de support composé des signaux des cinq meilleures prédictions de la sous-tâche d'identification.

Néanmoins, ce cadre sera d'une utilité limitée si le nombre de locuteurs à identifier devient trop élevé. Malgré cela, il est possible d'utiliser ces techniques dans un système d'ASR de bout en bout lorsque le vocabulaire est limité, comme dans l'expérience décrite dans [Eloff et al., 2019]. Ce cadre d'apprentissage a été utilisé dans la reconnaissance d'émotions [Feng and Chaspari, 2021]. Dans leurs expériences, ils ont utilisé leur approche sur une version modifiée d'IEMOCAP [Busso et al., 2008] utilisant une tâche à 3-ways (ce qui dénote des autres approches vues précédemment avec l'utilisation de quatre classes). Ainsi, ils ont obtenu un rappel moyen non pondéré de 67,4% avec une configuration à 10-shots. Ceci étant des résultats encourageants.

Les réseaux de correspondance

L'approche de Matching Networks décrit dans [Vinyals et al., 2016] est un cadre few-shot conçu pour être entraîné avec un ensemble d'épisodes multiples (avec typiquement 5 à 25-ways), qui consiste en un modèle unique φ. Ce modèle évalue les nouveaux exemples étant donné l'ensemble de support $\mathcal{S}$ comme dans le cadre siamois :

$$\varphi(\hat{x}, \mathcal{S}) :\rightarrow \hat{y} \tag{4.7}$$

Dans l'apprentissage par correspondance, φ se présente comme suit :

$$\varphi(\hat{x}, \mathcal{S}) = \sum_{(x_i, y_i) \in \mathcal{S}} a(\hat{x}, x_i) y_i \tag{4.8}$$

avec, a étant le noyau d'attention.

Dans [Vinyals et al., 2016] ce noyau d'attention est le suivant :

$$a(\hat{x}, x_i) = \text{softmax}(c(f(\hat{x}), g(x_i))) \tag{4.9}$$

où c est la distance en cosinus, f et g sont des fonctions d'intégration.

Vinyals et al. ont utilisé une architecture récurrente pour moduler la représentation de f avec l'ensemble de support S [Vinyals et al., 2016]. L'objectif est que f suive le même type de représentation que g. Pour ce faire, la fonction g est la suivante :

$$g(x_i) = \overrightarrow{h_i} + \overleftarrow{h_i} + g'(x_i) \tag{4.10}$$

où $\overrightarrow{h_i}$ et $\overleftarrow{h_i}$ représentent une sortie bi-LSTM sur $g'(x_i)$, qui est un DNN.

Ensuite, la fonction f est la suivante :

$$f(\hat{x}) = attLSTM(f'(\hat{x}), g(S), m) \tag{4.11}$$

avec, $attLSTM$ étant un LSTM nécessitant un nombre fixe de récurrences (ici m), $g(S)$ représentant l'application de g à chaque x_i de l'ensemble S. f' est un DNN ayant la même architecture que g', mais ne partageant pas nécessairement les valeurs des paramètres.

L'apprentissage de ce cadre consiste donc à maximiser le logarithme de la vraisemblance de $varphi$ étant donné les paramètres de g et f.

La figure 4.3 illustre la phase d'apprentissage d'un modèle de réseau de correspondance. Pour la phase de prédictions du modèle sur de nouveaux échantillons, $g(\mathcal{S})$ peut être précalculé pour gagner du temps de calcul. Néanmoins, les réseaux de correspondance présentent les mêmes inconvénients que les réseaux siamois lorsque q et/ou K deviennent trop élevés. En outre, l'ajout de nouvelles classes à un modèle de réseau de correspondance déjà entraîné n'est pas aussi facile que pour les modèles utilisant des réseaux siamois. En effet, cela nécessite de réentraîner le modèle de réseau de correspondances pour ajouter un élément à l'ensemble de support. Malgré ces inconvénients, l'apprentissage par appariement a montré de meilleurs résultats que le cadre siamois sur des ensembles de données d'images [Vinyals et al., 2016]. C'est pourquoi il devrait être étudié dans le traitement de la parole pour voir si c'est toujours le cas.

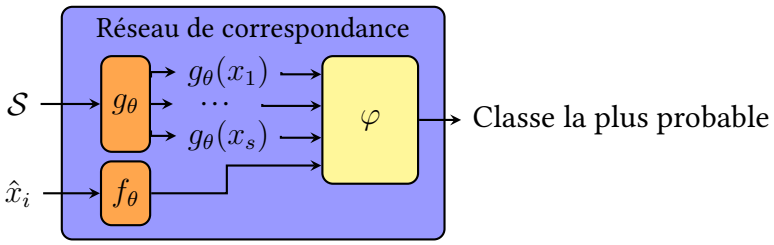


FIGURE 4.3 – Illustration d'un réseau de correspondances pour prédire la classe d'un nouvel exemple $\hat{x}_i$.

Réseaux Prototypiques

Les réseaux prototypiques (Prototypical Network [Snell et al., 2017]) sont conçus pour fonctionner avec de multiples épisodes. Pour les réseaux prototypiques, le modèle φ fait ses prédictions en fonction de l'ensemble de support $\mathcal{S}$ d'un épisode comme les approches précédemment vues. Ils utilisent les épisodes d'entraînement comme des mini-batchs pour obtenir le modèle final. Ce modèle est formulé comme suit :

$$\varphi(\hat{x}, S) = softmax_k(-d(f(\hat{x}), \mathbf{c}_k)) \tag{4.12}$$

où $\mathbf{c}_k$ est le prototype de la classe k, d étant une divergence de Bregman (pour leurs propriétés utiles en optimisation, voir [Snell et al., 2017] pour plus de détails), qui possède également la propriété suivante : $\mathbf{R}^n \times \mathbf{R}^n \to [0, +\inf[$.

Snell et al. ont utilisé la distance euclidienne pour d au lieu de la distance cosinus utilisée dans les articles sur le méta-apprentissage et l'apprentissage par correspondance [Snell et al., 2017]. En conséquence, ils obtiennent de meilleurs résultats dans leurs expériences. Ensuite, ils vont plus loin en réduisant la distance l'euclidienne à une fonction linéaire.

Dans le cadre prototypique, il n'y a qu'un seul prototype pour chaque classe k comme l'illustre la figure 4.4. Il est calculé comme suit :

$$\mathbf{c}_k = \frac{1}{|\mathcal{S}_k|} \sum_{(x_i, y_i) \in \mathcal{S}_k} f(x_i) \tag{4.13}$$

avec f étant une fonction de correspondance telle que $\mathbb{R}^D \rightarrow \mathbb{R}^M$ et $\mathcal{S}_k$ étant les échantillons avec k de l'ensemble de support.

Les réseaux prototypiques ne nécessitent qu'une seule comparaison par classe et non q par classe pour un apprentissage q-shot comme dans les réseaux d'apprentissage siamois et par correspondance. C'est pourquoi ce cadre est moins sujet au problème de calcul élevé pour la prédiction de nouveaux échantillons, car il n'est influencé que par un K élevé. Il sera certainement insuffisant pour les systèmes ASR de bout en bout sur la langue anglaise en raison des problèmes de grand vocabulaire décrits dans le sous-chapitre 4.2.1, mais c'est un pas vers cela.

Tandis qu'en reconnaissance du locuteur, les réseaux prototypiques ont été utilisés sur une portion de Voxceleb1 [Nagrani et al., 2017] par [Anand et al., 2019]. Les auteurs ont obtenu une précision de 72,77% dans un problème 20-ways avec 5-shots, ce qui est prometteur.

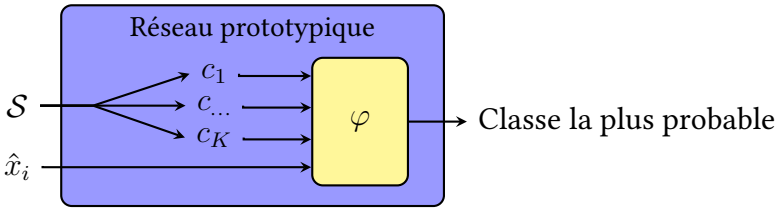


FIGURE 4.4 – Illustration d'un réseau prototypique pour prédire la classe d'un exemple $\hat{x}_i$.

Le méta-apprentissage

Les systèmes de méta-apprentissage (Meta-Learning en anglais) [Ravi and Larochelle, 2017] sont conçus pour être entraînés sur plusieurs épisodes (également appelés jeux de données). Dans cette approche, un modèle « apprenti » ($\mathcal{T}$) avec les paramètres $\theta^{\mathcal{T}}$ est entraîné dès le début de chaque épisode, généralement ce modèle a une architecture DNN classique. L'ensemble de support et l'ensemble de requêtes dans les épisodes sont considérés comme l'ensemble d'apprentissage et l'ensemble de tests pour le modèle d'apprentissage.

Parallèlement à ce modèle d'apprentissage, un second modèle est formé : le méta-modèle ($\mathcal{M}$) avec les paramètres $\theta^{\mathcal{M}}$. Ce méta-modèle est la clé du méta-apprentissage, il consiste à surveiller le modèle « apprenti » en actualisant les paramètres $\theta^{\mathcal{T}}$. Pour entraîner ce méta-modèle, Ravi et al. proposent d'échantillonner des épisodes iid (indépendantes et identiquement distribuées) de P pour former le méta-dataset ($\mathcal{D}$) [Ravi and Larochelle, 2017]. Ce méta-dataset de données est composé d'un ensemble d'apprentissage ($\mathcal{D}_{train}$), d'un ensemble de validation ($\mathcal{D}_{valid}$) et d'un ensemble de test ($\mathcal{D}_{test}$).

Pendant que le modèle « apprenti » s'entraîne sur un épisode $\mathcal{E}_j$, le méta-modèle est utilisé pour actualiser ses paramètres :

$$\theta_t^{\mathcal{T}_j} = \mathcal{M}(\theta_{t-1}^{\mathcal{T}_j}, \mathcal{L}^{\mathcal{T}_j}, \nabla_{\theta_{t-1}^{\mathcal{T}_j}} \mathcal{L}^{\mathcal{T}_j}) \tag{4.14}$$

avec $\mathcal{L}^{\mathcal{T}_j}$ étant la fonction de cout du modèle d'apprentissage appris avec l'épisode $\mathcal{E}_j$ et $\theta_{t-1}^{\mathcal{T}_j}$ étant les paramètres du modèle d'apprentissage à l'étape $t-1$. De plus, $\mathcal{M}$ doit deviner les poids initiaux des modèles « apprentis » à l'étape $t=0$ ($\theta_0^{\mathcal{T}_j}$).

La courbe d'apprentissage (fonction de coût) du modèle stagiaire avec $\mathcal{E}_j$ est vue dans [Ravi and Larochelle, 2017] comme une séquence qui peut être l'entrée du méta-modèle $\mathcal{M}$. Par souci de simplicité, nous utiliserons la notation de $\mathcal{T}$ au lieu de $\mathcal{T}_j$ pour les prochains paragraphes. La figure 4.5 illustre les étapes d'apprentissage du stagiaire à l'aide du méta-modèle.

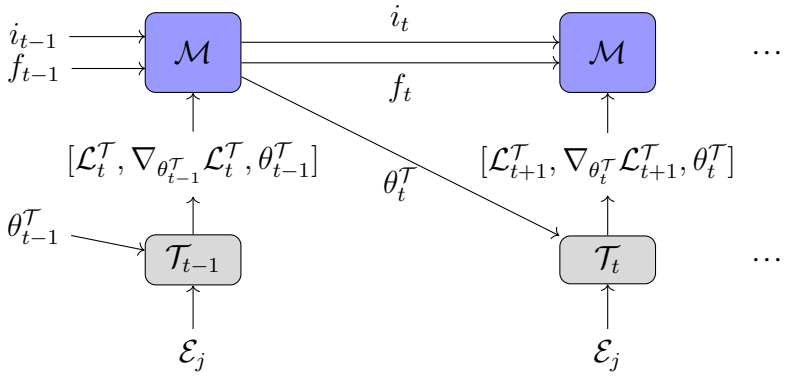


FIGURE 4.5 – Illustration du méta-apprentissage pour la formation avec l'épisode $\mathcal{E}_j$ à l'étape t. Ici, le méta-modèle $\mathcal{M}$ traite les différentes étapes d'apprentissage de « l'apprenti » $\mathcal{T}$ comme une séquence.

Mise à jour des paramètres de l'apprentissage Selon Ravi et Larochelle, le processus d'apprentissage de $\mathcal{T}$ utilisant la mise à jour feedforward classique sur l'épisode E_j est semblable à la porte de mise à jour c_t du cadre LSTM [Ravi and Larochelle, 2017]. Dans le cadre du méta-apprentissage, c_t est utilisé comme l'estimateur $\theta_t^{\mathcal{T}}$, comme suit :

$$\theta_t^{\mathcal{T}} = f_t \odot \theta_{t-1}^{\mathcal{T}} + i_t \odot \tilde{\theta_t^{\mathcal{T}}} \tag{4.15}$$

avec $\tilde{\theta_t^{\mathcal{T}}} = -\alpha_t \nabla_{\theta_{t-1}^{\mathcal{T}}} \mathcal{L}_t^{\mathcal{T}}$ étant le terme de mise à jour des paramètres $\theta_{t-1}^{\mathcal{T}}$, f_t étant la porte d'oubli et i_t la porte de mise à jour.

Paramètres du méta-modèle Les deux i_t et f_t font partie du méta-modèle d'apprentissage. Dans le cadre du méta-apprentissage, la porte de mise à jour est formulée comme suit :

$$i_t = \sigma(\mathbf{W}_I.[\nabla_{\theta_{t-1}^{\mathcal{T}}} \mathcal{L}_t^{\mathcal{T}}, \mathcal{L}_t^{\mathcal{T}}, \theta_{t-1}^{\mathcal{T}}, i_{t-1}] + \mathbf{b}_I) \tag{4.16}$$

avec $\mathbf{W}_I$ et $\mathbf{b}_I$ étant des paramètres de $\mathcal{M}$. La porte de mise à jour est utilisée pour contrôler le terme de mise à jour dans l'équation 4.15, comme le taux d'apprentissage dans l'approche feedforward classique.

Ensuite, la porte d'oubli dans le cadre du méta-apprentissage est formulée comme suit :

$$f_t = \sigma(\mathbf{W}_F.[\nabla_{\theta_{t-1}^{\mathcal{T}}} \mathcal{L}_t^{\mathcal{T}}, \mathcal{L}_t^{\mathcal{T}}, \theta_{t-1}^{\mathcal{T}}, f_{t-1}] + \mathbf{b}_F) \tag{4.17}$$

avec les paramètres $\mathbf{W}_F$ et $\mathbf{b}_F$ de $\mathcal{M}$. Cette porte permet de décider si l'apprentissage du réseau « apprenti » doit recommencer ou non. Cela peut être utile pour éviter le problème d'un minimum local sous-optimal. Notez que cette porte est absente dans les approches feedforward classiques (où cette porte est égale à un).

Le modèle de l'apprenti ($\mathcal{T}$) de cette approche peut être n'importe quel type de modèle, tel qu'un DNN. Il peut donc bénéficier des avantages d'autres approches. Il peut également éviter les inconvénients du réseau neuronal siamois, car il peut utiliser n'importe quelle autre approche (généralement un DNN classique). Ce cadre est intéressant pour former des modèles efficaces pour le traitement de la parole (en termes de vitesse d'apprentissage) lorsque nous avons plusieurs tâches d'ASR avec différents vocabulaires. Par exemple, supposons que nous ayons les types d'épisodes vocaux suivants : composition de numéros, commandes à un robot A et commandes à un robot B. Le modèle peut initialiser de bons filtres pour les premières couches (car cela implique toujours un traitement de la parole). Un autre exemple pourrait être l'entraînement de modèles acoustiques pour plusieurs langues (chaque épisode correspondant à une langue).

Réseau neuronal graphique

Les réseaux de neurones graphiques (GNN) sont utilisés par Garcia et Bruna pour approche en few-shot [Garcia and Bruna, 2018]. Cette approche est conçue pour être utilisée avec plusieurs épisodes, que les auteurs appellent des tâches. Dans ce cadre, un modèle est utilisé sur un graphe complet $G : G = (V, E)$ et chaque nœud correspond à un exemple. Pour l'apprentissage en quelques shots, un GNN consiste à appliquer des couches de convolution graphique sur le graphe G.

Les sommets initiaux pour deviner la vérité terrain d'une requête $\tilde{x_i}$ grâce à l'ensemble de requêtes $\mathcal{Q}$ sont construits comme suit :

$$\begin{aligned} V^{(0)} = ((&Enc(x_1), h(y_1)), \ldots, (Enc(x_s), h(y_s)), \\ &(Enc(\bar{x_1}), u), \ldots, (Enc(\bar{x_r}), u) \\ &(Enc(\tilde{x_i}), u)) \end{aligned} \tag{4.18}$$

où Enc est une fonction d'extraction de caractéristiques (un réseau de neurones ou toute autre technique classique d'extraction de paramètres), h la fonction d'encodage one-hot et $u = K^{-1}\mathbf{1}_K$ une distribution uniforme pour les exemples avec des étiquettes inconnues (les exemples non supervisés de $\bar{\mathbf{x}}$ et/ou de l'ensemble de requêtes $\mathcal{Q}$).

Les sommets de chaque couche l (0 étant les sommets initiaux) seront dorénavant notés :

$$V^{(l)} = (v_1, \ldots, v_n) \tag{4.19}$$

où $n = s + r + 1$ et $V^{(l)} \in \mathbb{R}^{n*d_l}$.

Chaque couche (avec une illustration d'une couche dans la Figure 4.6) dans un GNN est calculée comme suit :

$$V^{(l+1)} = Gc(V^{(l)}, A^{(l)}) \tag{4.20}$$

avec $A^{(l)}$ étant les opérateurs d'adjacence construits depuis $V^{(l)}$ et Gc étant une convolution du graphe.

Construction des opérateurs d'adjacence L'opérateur d'adjacence utilise l'ensemble :

$$A^{(l)} = \{\tilde{A}^{(l)}, \mathbf{1}\} \tag{4.21}$$

avec $\tilde{A}^{l)}$ étant la matrice d'adjacence de $V^{(l)}$.

Pour chaque $(i, j) \in E$ (rappelez-vous que nous avons des graphes complets), nous calculons les valeurs de la matrice d'adjacence comme suit :

$$\tilde{A}^{(l)}_{i,j} = \phi(v_i^{(l)}, v_j^{(l)}) \tag{4.22}$$

où :

$$\phi(v_i^{(l)}, v_j^{(l)}) = f(|v_i^{(l)} - v_j^{(l)}|) \tag{4.23}$$

avec f étant un perceptron multicouche dont les paramètres sont notés θ_f. On normalise ensuite $\tilde{A}^{(l)}$ grâce à la fonction softmax sur chaque ligne.

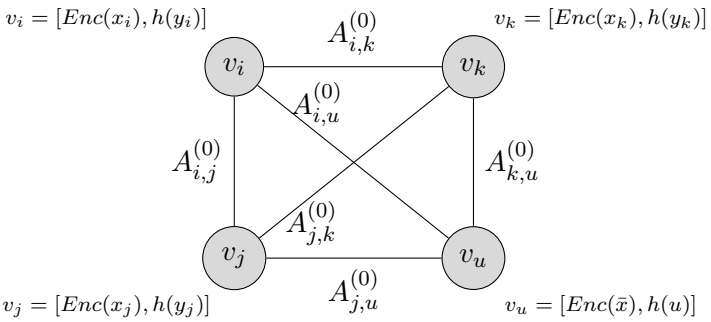


FIGURE 4.6 – Illustration de l'entrée de la première couche (ici une convolution graphique) d'un GNN. Ici, nous avons trois échantillons (représentés par les sommets v_i, v_j et v_k) dans l'ensemble de support et une requête (représentée par le sommet v_u).

Convolution des graphes La convolution du graphe nécessite la construction de tous les opérateurs d'adjacence et est calculée comme suit :

$$Gc(V^{(l)}, A^{(l)}) = \rho(\sum_{B \in A} BV^{(l)}\theta^{(k)}_{B,l}) \tag{4.24}$$

avec B étant un opérateur d'adjacence de A, $\theta^{(k)}_{B,l} \in \mathbb{R}^{d_{l-1}, d_l}$ paramètres apprenables et ρ étant une linéarité ponctuelle (généralement leaky ReLU).

Formation du modèle La sortie du modèle GNN résultant est un mappage des sommets vers un complexe K qui donne la probabilité que $\tilde{x_i}$ soit dans la classe k. V. Garcia et J. Bruna ont utilisé l'entropie croisée pour entraîner le modèle en utilisant tous les autres échantillons de l'ensemble de requêtes $\mathcal{Q}$ [Garcia and Bruna, 2018]. Par conséquent, le cadre GNN few-shot consiste à apprendre les paramètres θ_f et $\theta_{1,l} \ldots \theta_{card(A),l}$ avec tous les épisodes.

GNN sur l'audio Cette approche a été utilisée par [Zhang et al., 2019] sur des problèmes de classification audio à cinq classes. Les épisodes à cinq classes sont sélectionnés aléatoirement dans l'ensemble de données initial : AudioSet [Gemmeke et al., 2017] pour créer les épisodes d'entraînement à cinq classes et les données des programmes TV (de [Zhang et al., 2006]) pour créer les épisodes de test à cinq classes.

Zhang et al. comparent l'utilisation de l'attention par classe (ou intra-classe) et de l'attention globale, qui a donné les meilleurs résultats [Zhang et al., 2019]. Ils l'ont appliquée pour chaque couche. Leurs expériences ont été réalisées pour 1-shot, 5-shots et 10-shots avec une précision respective de $69,4\% \pm 0,66$, $78,3\% \pm 0,46$ et $83,6\% \pm 0,98$. De tels résultats sont un encouragement pour l'utilisation de l'apprentissage à quelques instants pour les signaux vocaux. Néanmoins, ce cadre ne permet pas l'utilisation de classes et de plans multiples par épisode, ce qui augmente le nombre de nœuds et donc les calculs en temps différé. Par conséquent, il n'est pas adapté aux problèmes de grand vocabulaire.

4.3 Expérience few-shot sur corpus cancer

Lorsque j'ai réalisé mes expériences, je n'avais pas à disposition de bases de données semblables à celle du corpus cancer. C'est pourquoi nous n'avons pas choisi de méthodes liées au méta-apprentissage. En perspective, il sera intéressant d'essayer de telles techniques avec les futurs corpora disponibles au travers de la parolothèque pour lesquels un corpus de patient atteint de la maladie de Parkinson [Ghio et al., 2012] (au travers du projet RUGBI) et d'un autre corpus de maladies cancéreuses buccales (au travers du projet DAPADAF-E [19]). De plus, d'autres corpora tels que celui du projet MonPaGe [Pernon et al., 2020] nous semblent pertinents pour utiliser de telles approches.

Pour la suite de nos expériences, nous avons choisi d'utiliser des techniques de few-shot au niveau phonémique. Le choix du niveau phonémique est lié à la revue systématique [Balaguer et al., 2020] qui ont démontré que le niveau phonémique est le niveau le plus souvent utilisé pour l'analyse de mesures d'intelligibilité de production.

Ainsi, nous considérons que les techniques de réseau siamois et prototypique sont les plus adaptées pour une telle tâche. Néanmoins, pour ne pas invalider les scores perceptifs, nous n'avons pas utilisé des méthodes d'augmentation de données (bien que ces méthodes soient utilisées en image pour les réseaux siamois et prototypiques). C'est pourquoi nous avons décidé de tester ces techniques de few-shot avec une tâche de reconnaissance de phonème pour tester la viabilité d'une telle approche.

Nous commencerons par tester ces approches ainsi que la sélection de cette architecture en sous-chapitre 4.3.1 pour ensuite tester les meilleures architectures qui en ressortent en sous-chapitre 4.3.2.

4.3.1 Few-shot pour de la reconnaissance de phonème

Pour la reconnaissance de phonème, nous avons choisi le corpus TIMIT [Garofolo et al., 1993] car c'est une référence en traitement de la parole. Pour le choix du modèle, nous avions

19. https://www.irit.fr/SAMOVA/site/projects/dapadaf-e/

choisi l'état de l'art avec le code de disponible [20] et nous nous sommes tournés vers pytorch-kaldi [Ravanelli et al., 2019].

L'architecture du code de ce projet n'étant pas facilement manipulable pour réaliser du few-shot, une première étape consista à reproduire le modèle. L'avantage d'avoir eu le code original de pytorch-kaldi à permis d'utiliser les mêmes hyperparamètres que dans leur code.

Pour des raisons de rapidité d'implémentation, certaines simplifications ont été faites. Comme le non-emploi de liGRU (ici, remplacé par des GRU) et des fMLLR (ici, remplacé par des MFCC avec les deltas d'ordre 1 et 2). Enfin, nous avons utilisé la vérité terrain pour extraire les trames correspondant à chaque phonème. Ainsi, nous avons reconstruit l'architecture détaillée en table 4.6 pour la reconnaissance de 39 phonèmes.

Il est à noter que nous n'avons pas utilisé de mécanismes d'augmentation de données comme les auteurs de pytorch-kaldi, car nous ne pourrons pas les utiliser dans nos expériences avec les données pathologiques. Pour évaluer nos résultats, nous avons utilisé la précision phonémique (où le taux d'erreur phonémique $= 1 -$ précision phonémique) pour correspondre à la métrique habituelle utilisée dans l'apprentissage few-shot. Le meilleur résultat obtenu étant celui fondé sur les MFCC et nous a donné une précision d'environ 65% sur l'ensemble de tests de TIMIT. Ainsi, ce score est le maximum que nous espérions obtenir en implémentant les réseaux siamois et prototypiques.

Table 4.6 – Architecture du réseau de pytorch-kaldi utilisée.

N° couche	Type de couche	Paramètres
0	donnée en entrée	MFCC avec un fenêtrage de 25ms et un saut de 10ms
1	GRU bidirectionnels empilés	5 GRU de 550 cellules chacun
2	dropout	de 0,2
3	normalisation du batch	pour chaque direction
4	couche linéaire	128 filtres
5	normalisation du batch résultant	32 filtres
6	activation	leaky ReLU de paramètre 0,1
7	couche linéaire	39 filtres pour les 39 classes

Résultats few-shot pour 39 classes

Suite à ce premier résultat, nous avons tenté d'implémenter les solutions few-shot pour la reconnaissance de phonème sur TIMIT. Nos premières tentatives réutilisaient les couches 0 à 5 du tableau 4.6. Néanmoins, nous avons eu des difficultés à faire apprendre ces modèles (les problèmes étant similaires pour les réseaux siamois et prototypique). Dans nos premières expériences, les deux architectures convergeaient vers une diminution de la fonction de coût, mais cela engendrait une diminution du score de précision phonémique sur l'ensemble de tests (qui est le comportement non désiré). Pour résoudre ce problème, nous avons dû réaliser les modifications suivantes :

20. Selon https://paperswithcode.com/sota/speech-recognition-on-timit

1. Lors de l'apprentissage du modèle, nous avons présenté autant de paires positives que négatives pour chaque batch (spécifique aux réseaux siamois)
2. Nous avons fait en sorte d'équilibrer les classes présentes pour chaque batch (réseaux siamois et prototypiques).
3. L'équilibre des hommes et femmes pour les exemples de références (réseaux siamois et prototypiques).
4. Nous avons observé que la réduction du nombre de paramètres de l'architecture choisie améliorait les résultats. La nouvelle architecture utilisée se trouve en table 4.7. Ainsi, nous passons d'une architecture d'environ 23 millions de paramètres à environ 5 millions de paramètres. Par conséquent, nous supposons que pour notre cas (sans augmentation de données sur de la reconnaissance de phonèmes) ces méthodes ne permettent pas d'utiliser des modèles avec un grand nombre de paramètres.
5. Le choix des exemples de références importe pour environ 5% de précision dans nos expériences. Ainsi, le choix des exemples provenant de régions dialectales différentes [21] permet une augmentation des scores, mais n'est pas toujours possible en fonction des données à disposition.

TABLE 4.7 – Architecture utilisée pour les réseaux siamois et prototypique.

N° couche	Type de couche	Paramètres
0	Donnée en entrée	MFCC avec un fenêtrage de 25ms et un saut de 10ms
1	GRU bidirectionnels empilés	5 GRU de 256 cellules chacun
2	Dropout	de 0,2
3	Normalisation du batch	pour chaque direction
4	Couche linéaire	128 filtres

Ainsi pour la reconnaissance des 39 phonèmes anglais de TIMIT nous avons obtenu au mieux 32,58% pour le réseau siamois et 41,38% pour le réseau prototypique. Ces premiers résultats sont encourageants, surtout en considérant que nous avons utilisé 1,1% des données de l'ensemble d'entrainement de TIMIT (soit environ 15 minutes de signal audio). Pour comparer ces deux méthodes, nous avons réalisé deux tables de compilations (Table 4.8 et Table 4.9). Il est à noter que pour l'architecture des réseaux siamois, le temps de calcul de mon implémentation était conséquent (environ une journée pour une expérience) comparé à mon implémentation des réseaux prototypique (environ 6h pour une expérience). Cette différence est due à la nature des architectures et explique que le nombre d'essais est moins conséquent pour les réseaux siamois. En effet, dans les réseaux siamois, chaque référence est comparée à un nouvel échantillon, tandis que dans les réseaux prototypiques, seuls les prototypes sont comparés aux nouveaux échantillons. Par ces premiers résultats, nous considérons que l'architecture prototypique est plus intéressante comme les temps de calcul sont moins élevés et

21. Dans le jeu de données TIMIT, les données proviennent de huit régions dialectales des États-Unis. Ces dialectes ont pour origines : la Nouvelle-Angleterre, Nord, Sud, Nord du Midland, Sud du Midland, New York, Ouest et l'armée (personnes se déplaçant). Remarque : les auteurs ne peuvent pas garantir les limites des dialectes pour les deux dernières origines.

que les résultats sont meilleurs. De plus, on remarque que le réseau prototypique semble plus stable si le nombre d'exemples supervisés augmente (comparé au réseau siamois).

TABLE 4.8 – Résultats sur les 39 phonèmes de TIMIT avec le réseau siamois.

#shots	Précision sur test
10	22,76
25	32,58
40	27,90

TABLE 4.9 – Résultats sur les 39 phonèmes de TIMIT avec le réseau prototypique.

#shots	#queries	Précision sur test
10	15	39,76
15	15	37,82
10	20	41,16
20	15	41,12
15	20	41,33
20	20	41,38

Résultats few-shot pour 4 classes

Pour se rapprocher de notre objectif de classer les participants en quatre classes (gravement, moyennement, faiblement et non atteint par la maladie), nous avons choisi de créer une tâche de définition de traits phonémique. Ainsi, nous avons créé quatre classes que sont les voyelles, les fricatives, les nasales et les stops [22].

Pour cette expérience, nous avons utilisé exactement la même architecture que pour la reconnaissance des 39 phonèmes. Ainsi, nous obtenons un score de 71,6% de précision sur l'ensemble de tests pour le réseau siamois et un meilleur score de 71,31% pour le réseau prototypique. Il est à noter que pour cette expérience, nous utilisons au maximum (pour 40 shots) 160 exemples pour l'ensemble d'entraînement soit 0,1% de l'ensemble d'entraînement de TIMIT (c'est à dire environ 3 minutes de signal audio). Il est à noter qu'ici le réseau siamois obtient le meilleur résultat et que ces résultats sont toujours effectués sans augmentation de données. Un détail des résultats se trouve en Table 4.10 et Table 4.11. Ici, les deux réseaux obtiennent de moins bons résultats avec 40 exemples supervisés. Ainsi, cette expérience nous indique que ces deux architectures permettent d'obtenir des résultats semblables. Au vu des temps de calculs et de nos résultats, nous nous sommes tournés vers les réseaux prototypiques pour la suite de nos expériences avec de la parole pathologique.

22. La table de correspondance que nous avons utilisée pour transformer les phonèmes initiaux de TIMIT en ces classes se trouve ici : https://github.com/vroger11/audio_loader/blob/master/audio_loader/ground_truth/timit_map.csv

TABLE 4.10 – Résultats sur les quatre classes phonémiques de TIMIT avec le réseau siamois.

#shots	Précision sur test
10	55,84
20	64,64
25	55,66
30	71,60
40	69,02

TABLE 4.11 – Résultats sur les quatre classes phonémiques de TIMIT avec le réseau prototypique.

#shots	#queries	Précision sur test
10	15	63,93
10	20	66,10
15	15	66,48
15	20	66,19
20	15	71,31
20	20	64,79

4.3.2 Essais few-shot sur le corpus cancer

Suite à nos expériences sur TIMIT, nous avons décidé de tester le réseau prototypique sur notre corpus cancer. Pour ne pas trop nous écarter de ces expériences et pour avoir suffisamment d'exemples pour l'apprentissage de notre modèle, nous avons choisi de réaliser une tâche de classification de sévérité de la maladie par phonème.

Dans nos expériences, nous avons commencé par deux classes (voir les sous-chapitres 4.3.2 et 4.3.2) et finis par un découpage en trois classes (voir le sous-chapitre 4.3.2). Dans chaque cas, nous avons réalisé une tâche de classification pour chaque phonème bien que nous disposons d'une évaluation par fichier. Même si le score global du jury ne peut se répercuter sur l'ensemble de chaque fichier (certains phonèmes sont plus intelligibles que d'autres par un même participant) nous pensons que l'erreur moyenne sera suffisante pour prédire un score global par fichier à partir de plusieurs prononciations. Ainsi, la tâche choisie est une tâche intermédiaire pour réaliser notre objectif de prédire la tranche de score pour chaque participant (réalisé grâce à un vote majoritaire).

Pour récupérer les segments correspondant à chaque phonème, nous avons utilisé un alignement automatique forcé [23]. Ces segments sont donnés en entrée de notre modèle pour prédire la classe de sévérité. Cet alignement n'étant pas parfait, on ne dispose pas de tous les phonèmes prononcés par les patients. Pour l'ensemble de validation, nous avons utilisé 10 exemples par classe. Ceci pour laisser un maximum d'exemples pour l'ensemble de tests (au détriment d'un ensemble de validation moins représentatif du problème). Chaque participant se retrouve dans un ensemble uniquement (soit l'ensemble d'entrainement, de validation ou de tests) pour éviter les biais de surapprentissages.

23. Réalisé par Corinne Fredouille du laboratoire du LIA (https://lia.univ-avignon.fr/)

Problème à deux classes - premiers résultats

Pour le problème à deux classes, nous avons arbitrairement choisi d'opposer les personnes sévèrement atteintes par la maladie (avec un score perceptif de sévérité <6) des personnes pas ou peu atteinte par la maladie (sévérité >9,5). Ce type de problème peut s'avérer pertinent dans un milieu médical. De plus, cette expérience nous aide à tester la formalisation du problème et le réseau prototypique choisi.

Pour la suite, nous avons réalisé trois expériences avec notre modèle établi sur les MFCC ou les Mel spectrogrammes en entrée. Nous poursuivrons et en utilisant seulement les hommes de notre corpus pour l'apprentissage.

MFCC En utilisant les MFCC, notre modèle obtient une précision globale sur l'ensemble de test de 57,70% pour chaque phonème. Ainsi, nous avons pu créer une table de précision par phonèmes que l'on retrouve en Table 4.12. Grâce à cette table, on constate que les voyelles sont globalement mieux prédites que les autres phonèmes. De plus, on s'aperçoit que malgré le peu d'exemples présents dans l'ensemble de validation, nos scores de validation correspondent globalement aux résultats sur l'ensemble de tests.

TABLE 4.12 – Résultats de classification par phonème sur le corpus cancer du réseau prototypique pour le problème à deux classes en utilisant les MFCC.

Phonème	Précision sur test	Précision sur validation
eu	75,97	90
aa	66,47	70
ai	62,70	80
ei	62,45	80
au	61,82	80
nn	58,67	80
rr	58,13	70
in	58,1	70
bb	57,14	40
mm	56,71	60
ou	55,96	80
vv	55,14	70
oe	54,76	70
ii	54,69	60
yy	53,97	50
an	53,74	60
ll	53,29	70
ss	52,66	70
dd	51,90	40
kk	51,76	70
tt	43,97	50
pp	38,32	60

Ainsi pour la suite, pour améliorer les résultats de classification par participant, nous avons utilisé ces prédictions par phonèmes pour réaliser un vote majoritaire sur tous les phonèmes (sans sélection). Ensuite, nous avons pris le top 5 des phonèmes de l'ensemble de validation (après sélection) pour prédire un vote majoritaire sur ces phonèmes uniquement [24]. La table 4.13 regroupe l'ensemble de ces résultats. On remarque que la sélection des meilleurs phonèmes permet d'augmenter grandement la précision de notre système. De plus, une précision globale de phonèmes de 57,7% nous a permise d'atteindre des scores raisonnables.

TABLE 4.13 – Résultats sur le corpus cancer du réseau prototypique pour le problème à deux classes en utilisant les MFCC.

	Prédiction pour les femmes (8 fichiers)	Prédiction pour les hommes (13 fichiers)	Prédiction sur tous (21 fichiers)
Sans sélection	87,50%	69,23%	76,19%
Après sélection	100%	92,31%	95,24%

Mel spectrogramme Pour observer l'impact de l'entrée choisi par notre réseau, nous avons également utilisé les Mel spectrogrammes en entrées de notre modèle. Pour ce faire, nous avons utilisé le même fenêtrage que pour l'utilisation de MFCC. Ainsi, nous avons obtenu au mieux une précision globale par phonème sur l'ensemble de tests de 54,30%. Le détail par phonème se trouve en table 4.14. Ainsi comme pour l'utilisation des MFCC, on observe que notre modèle a plus de facilités avec les voyelles comparées aux autres phonèmes.

Comme pour les MFCC, nous avons réalisé des prédictions par fichier (à l'aide d'un vote majoritaire) sans et avec sélection de meilleurs phonèmes (que l'on retrouve en table 4.15). Nous avons retrouvé les mêmes tendances qu'avec l'utilisation des MFCC. Néanmoins, nous obtenons globalement de moins bons résultats suite à la sélection de meilleurs phonèmes. Ceci peut s'expliquer par l'optimisation de notre modèle pour les MFCC. Ainsi, nous avons décidé de nous concentrer sur les MFCC pour la suite de nos expériences.

Apprentissage sur les hommes Suite à ces expériences, nous souhaitions observer l'impact de l'équilibre homme/femme sur l'apprentissage de nos données. Ainsi, nous avons essayé d'utiliser seulement les hommes dans l'apprentissage (nous n'avons pas suffisamment de participantes pour essayer un apprentissage avec des femmes seulement). Comme nous disposons de moins de données pour l'apprentissage, nous avons pu réaliser l'apprentissage sur les phonèmes suivants : 'aa', 'ai', 'dd', 'ei', 'eu', 'll', 'rr', 'tt'. Ainsi, nous avons utilisé toutes les femmes dans l'ensemble de tests et certains hommes ont également été rajoutés en test (comme on utilise moins de phonèmes pour l'entraînement).

Malgré la difficulté en apparence accrue, nous obtenons une précision globale par phonèmes de l'ensemble de tests de 60,16%. On retrouve en table 4.16 la précision obtenue par phonème et comme pour les expériences précédentes, on retrouve les voyelles en phonème les mieux appris. Ainsi, malgré l'utilisation de moins de données et bien qu'aucune femme ne soit

24. Dans le cas où plusieurs phonèmes ont le même score nous prenons tous les phonèmes correspondants. Ainsi pour l'expérience avec les MFCC, nous avons pris les phonèmes suivants : ['nn', 'ai', 'ou', 'au', 'eu', 'ei']

TABLE 4.14 – Résultats sur le corpus cancer du réseau prototypique pour le problème à deux classes en utilisant les Mel spectrogrammes.

Phonème	Précision sur test	Précision sur validation
eu	75,32	100
au	65,45	80
aa	64,69	70
ai	62,70	90
ou	60,55	90
ei	59,29	90
rr	58,68	70
oe	58,33	70
nn	57,33	70
bb	55,56	40
ii	54,69	70
ss	54,44	80
dd	54,29	60
vv	54,21	70
yy	53,97	60
an	53,74	70
ll	53,06	80
kk	51,76	70
mm	46,95	60
tt	44,36	60
in	39,05	50
pp	38,32	60

TABLE 4.15 – Résultats sur le corpus cancer du réseau prototypique pour le problème à deux classes en utilisant les Mel spectrogrammes.

	Prédiction pour les femmes (8 fichiers)	Prédiction pour les hommes (13 fichiers)	Prédiction sur tous (21 fichiers)
Sans sélection	87,5%	76,92%	80,95%
Après sélection	75%	84,62%	80,95%

utilisée lors de l'apprentissage, nous obtenons de meilleurs résultats sur un ensemble de tests plus important (la table 4.17 correspond aux prédictions de fichiers de l'ensemble de tests). Le plus notable est que l'on obtient une précision d'environ 96% sur l'ensemble des femmes alors que nous utilisons seulement des hommes à l'apprentissage. Ainsi, le modèle semble capable de généraliser sur le sexe.

Ces meilleurs résultats peuvent selon nous s'expliquer par l'ensemble de phonèmes utilisé pour l'apprentissage. Dans l'extrait de la chèvre de monsieur Seguin, certains phonèmes sont prononcés plusieurs fois dont principalement les voyelles. Ainsi le modèle a en entrée ces phonèmes dans les contextes variés (provenant de mots différents), ce qui permettrait au modèle

de différencier plus facilement les sons sévèrement atteints par la maladie des sons peu ou pas atteints. Ainsi pour vérifier cette hypothèse, nous avons créé un résultat affiné utilisant seulement les phonèmes répétés pour l'apprentissage.

TABLE 4.16 – Résultats de classification par phonème sur le corpus cancer du réseau prototypique pour le problème à deux classes en utilisant les MFCC avec apprentissage sur les hommes uniquement.

Phonème	Précision sur test	Précision sur validation
eu	79,75	90
aa	69,69	80
ai	68,97	90
ei	68,77	80
rr	52,04	80
dd	49,89	70
tt	49,07	70
ll	44,62	50

TABLE 4.17 – Résultats sur le corpus cancer du réseau prototypique pour le problème à deux classes en utilisant les MFCC avec apprentissage sur les hommes uniquement.

	Prédiction pour les femmes (26 fichiers)	Prédiction pour les hommes (18 fichiers)	Prédiction sur tous (44 fichiers)
Sans sélection	92,31%	83,33%	88,64%
Après sélection	96,15%	88,89%	93,18%

Problème à deux classes - résultats affinés

Suite à nos précédentes expériences, nous avons reproduit notre expérience avec seulement les phonèmes prononcés plusieurs fois (et reconnus par l'alignement automatique) dans l'extrait de la lettre de monsieur Seguin prononcé par les participants. Ainsi, nous obtenons une précision globale par phonème de 64,03%.

Malgré tout, comme nous pouvons le constater dans la table 4.18 que les voyelles semblent toujours obtenir de meilleurs résultats. On pourrait supposer qu'elles sont plus porteuses d'intelligibilité que les autres phonèmes (pour les maladies atteintes sur nos patients), mais d'autres expériences seraient nécessaires pour vérifier ceci. Néanmoins, les phonèmes moins bien reconnus semblent décisifs pour la décision par fichier vu que les résultats sans sélection sont moins bons (voir la table 4.19). Ainsi pour les expériences suivantes, nous utiliserons la même liste de phonèmes pour l'apprentissage de notre modèle.

Problème à trois classes

Pour séparer notre base de données en trois classes, nous avons utilisé toutes nos données à disposition (comparé à nos expériences précédentes). Néanmoins, comme il n'existe pas à notre

TABLE 4.18 – Résultats de classification par phonème sur le corpus cancer du réseau prototypique pour le problème à trois classes en utilisant les MFCC.

Phonème	Précision sur test	Précision sur validation
eu	80,23	80
ai	68,16	80
aa	68,00	90
ei	67,46	60
rr	60,23	70
ll	59,83	80
dd	53,87	40
tt	50,41	60

TABLE 4.19 – Résultats sur le corpus cancer du réseau prototypique pour le problème à deux classes en utilisant les MFCC et les phonèmes répétés.

	Prédiction pour les femmes (19 fichiers)	Prédiction pour les hommes (21 fichiers)	Prédiction sur tous (40 fichiers)
Sans sélection	100%	80,95%	90%
Après sélection	89,47%	80,95%	85%

connaissance de consensus sur le choix des bornes, nous avons testé avec plusieurs bornes. La table 4.20 résume nos résultats avec des bornes différentes. Ainsi, nous obtenons des résultats mitigés. Cela peut être dû au choix de frontières pouvant être trop brutales (il nous est difficile de différencier des participants ayant obtenu des scores similaires [25])

Néanmoins, notre modèle fait peu d'erreurs entre les classes extrêmes (voir la matrice de confusion en table 4.21). Environ 2% d'erreurs sont non souhaitées (de gravement atteint prédit en très intelligible). Ces résultats sont encourageants, mais nécessitent d'être approfondis pour être utilisables en milieu médical. Il est à noter que nous utilisons 23 fichiers pour l'entrainement (ensemble de query + ensemble d'ancres + ensemble de validation) et 91 fichiers pour le test. Ainsi, une approche par réseau prototypique montre un potentiel pour les approches few-shot. Néanmoins, l'utilisation d'autres approches nécessitera l'adaptation de modèles différents (comme pour les réseaux de graphes) et demandera ainsi plus de temps de conception et d'apprentissage.

4.4 Résumé et discussion

Dans ce chapitre, nous avons commencé par les systèmes de traitement de la parole les plus avancés. Ces systèmes requièrent une grande quantité de données (de l'ordre du millier d'heures) et ne sont pas adaptés aux problèmes d'utilisation de la parole qui ne disposent pas de ressources suffisantes. Nous avons également étudié les techniques nécessitant des corpus

25. Nous supposons que ce problème est identique pour notre modèle.

TABLE 4.20 – Résultats sur le corpus cancer du réseau prototypique pour le problème à trois classes en utilisant les MFCC et les phonèmes répétés.

Borne inférieure	Borne supérieure	Sans sélection	Après sélection
6	9,5	42,39%	46,74%
5	9,5	30%	33,33%
6	9	47,31%	53,76%
5	9	54,95%	42,86%
6	8,5	42,55%	42,55%
5	8,5	42,39%	41,30%

TABLE 4.21 – Matrice de confusion sur le corpus cancer du réseau prototypique pour le problème à trois classes en utilisant les MFCC et les phonèmes répétés.

	Scores bas	Scores moyens	Scores élevés
Scores bas	11	23	2
Scores moyens	1	22	11
Scores élevés	0	4	17

moins volumineux grâce à l'augmentation des données, la transposition de domaine, les modèles nécessitant moins de paramètres, l'approche multi-tâches et l'apprentissage par transfert. Néanmoins, certaines techniques telles que l'augmentation de données et la transposition de domaine ne sont pas envisageables pour notre tâche (car elles peuvent modifier les signaux initiaux de telle façon que les scores perceptifs soient invalidés). Suite à nos premières expériences et au travers des corpus utilisés par la littérature, ces techniques sont moins efficaces dans un contexte de données limitées (comme pour notre corpus).

C'est pourquoi nous nous sommes tournés sur un état de l'art de techniques initialement utilisées en image pour des corpus contenant un faible nombre de données : les techniques de "few-shot". Le principal inconvénient des techniques examinées est la quantité de calculs nécessaires pour les grands ensembles de données (comme LibriSpeech de [Panayotov et al., 2015]) face aux modèles SOTA que nous avons examinés dans le sous-chapitre 4.1.2. Néanmoins, nous avons considéré certains travaux récents qui utilisent déjà les techniques few-shot sur la parole avec des résultats prometteurs. De telles techniques semblent utiles pour les tâches classiques de parole sur des locuteurs atteints de pathologies cancéreuses. Acquérir une grande quantité de données est long et laborieux pour certains patients (avec des pathologies sévères). Nous pensons que les techniques de "few-shot" peuvent aider la communauté à résoudre ce problème. Nos premiers résultats sur le jeu de données TIMIT indiquent qu'une telle approche est adaptable à la reconnaissance de phonèmes. En effet, nous avons réussi à obtenir une précision de 41% avec seulement 1% des données disponibles pour l'apprentissage du corpus TIMIT. De plus, ce résultat a été obtenu sans recourir à l'augmentation de données (ce qui devrait améliorer ce premier résultat). Enfin, avec encore moins de données, nous obtenons une précision supérieure à 71% sur la classification de quatre classes phonétiques.

Suite à ces premiers résultats, nous avons utilisé le réseau prototypique pour apprendre des classes établies sur l'index de sévérité de la parole. Avec cette approche, nous avons identifié

que les voyelles semblent plus simples à apprendre (sur notre corpus cancer) que les autres phonèmes. Ainsi, le score des experts peut potentiellement être influencé par la prononciation des voyelles des participants. Ceci peut s'expliquer du fait que le corpus est majoritairement constitué de voyelles et demande une étude plus approfondie pour être confirmée. Excepté ceci (qui est lié à notre formalisation du problème d'apprentissage), nous obtenons un résultat similaire avec ce que nous avons pu voir avec l'approche de Fang et al. [Fang et al., 2017] dans le chapitre 2. Néanmoins, bien que nos résultats soient insuffisants, des axes d'améliorations sont possibles. Ainsi, nous pouvons envisager l'utilisation d'augmentations ne dénaturant pas excessivement le son des patients. Cependant, cela nécessiterait une étude auprès des experts de notre corpus pour identifier si leurs scores restent inchangés pour chaque augmentation utilisée.

Notre corpus étant limité, la nécessité d'utiliser une partie de nos données pour l'apprentissage d'un modèle est un handicap. Ainsi, l'utilisation de corpus externes et de techniques non-supervisées est au cœur du chapitre suivant.

5

Approches non-supervisées et l'utilisation de corpus externes

Alors que les méthodes supervisées sont prometteuses, elles ne nous ont pas aidé à réaliser des régressions de nos scores avec des résultats satisfaisants. Ainsi, nous nous sommes tournés vers des méthodes non supervisées qui ont pour principe d'éviter d'utiliser la vérité terrain pour modéliser les données. Nous nous sommes également tournés sur l'utilisation de corpus externes pour éviter d'utiliser une partie de nos données pour l'apprentissage d'un modèle. Ainsi, pour ce travail, je me suis concentré sur des méthodes de projection (sous-chapitre 5.1) et j'ai proposé un nouveau cadre d'applications pour créer un score entropique de la parole (sous-chapitre 5.2).

5.1 Projection des données

Nous avons utilisé le détecteur d'activité vocale vu dans le sous-chapitre 3.3.2, puis extrait 13 MFCC sur des fenêtres de 128 ms et de recouvrement d'un demi et chaque dimension a été normalisé [26] par fichier. Chaque fenêtre d'analyse a été donnée en entrée d'une projection en 2D par t-SNE [Van der Maaten and Hinton, 2008] [27]. La t-SNE est une méthode non-linéaire permettant de projeter des données d'un espace à grande dimension en un espace avec moins de dimensions (ici dans nos expériences, nous avons choisi deux dimensions). Cet algorithme essaie de conserver la distance entre chaque point de l'espace original sur l'espace projeté (qui est de plus faible dimension). Cette projection des données a été réalisée pour chacune de nos tâches : lecture, description d'image et /a/ tenus.

Nous avons coloré ces projections avec les scores (arrondis à la valeur la plus proche) d'intelligibilité et de sévérité : voir résultats sur les figures 5.1, 5.2 et 5.3. À noter que pour les /a/ tenus, nous avons utilisé les scores de sévérité et d'intelligibilité obtenus sur la tâche de description d'image.

Le résultat sur la tâche de /a/ tenu nous montre que les productions des participants se regroupent en utilisant les MFCC. Certains regroupements sont corrélés au score perceptif de

26. Pour cette expérience, nous avons utilisé la norme l_{max}.

27. Nous avons également essayé par analyse en composantes principales sans pouvoir séparer visuellement les données

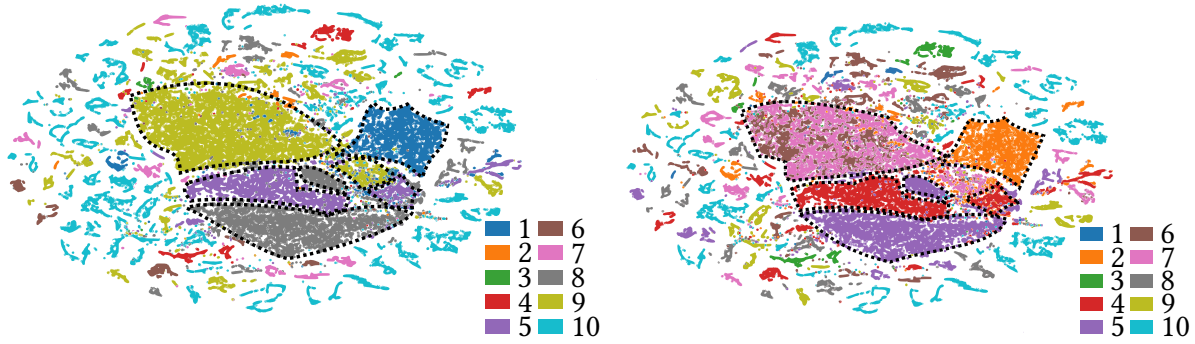


FIGURE 5.1 – Projection t-SNE en deux dimensions des fenêtres MFCC de parole sur la tâche de /a/ tenu. À gauche, la coloration correspond à l'intelligibilité et à droite, la coloration correspond à la sévérité (scores perceptifs de la tâche de description d'image). Nous avons détouré en pointillés noirs les différentes zones d'intérêts de ces projections.

sévérité et d'intelligibilité obtenu par les participants. Justement, sur la projection colorée avec l'intelligibilité, nous pourrions aisément séparer les amas correspondants aux scores 9, 5, 8 et 1. Nous pourrions en faire de même avec la sévérité, même si les frontières ne sont pas aussi bien définies. Ce résultat corrobore avec les résultats obtenus pour nos approches supervisées en few-shot, c.-à-d. où les voyelles semblaient plus simples pour l'apprentissage du concept de sévérité (ou d'intelligibilité). Néanmoins, il faudrait disposer de la prononciation isolée de tous les phonèmes français pour vérifier si la séparation des voyelles corrèle davantage aux scores d'intelligibilité et de sévérité que la projection de consonnes.

Cependant, nous ne retrouvons pas cet effet sur les projections des tâches de lecture et de description d'image. Ce qui est sûrement dû au nombre de phonèmes variés présents ou à la quantité de données à projeter.

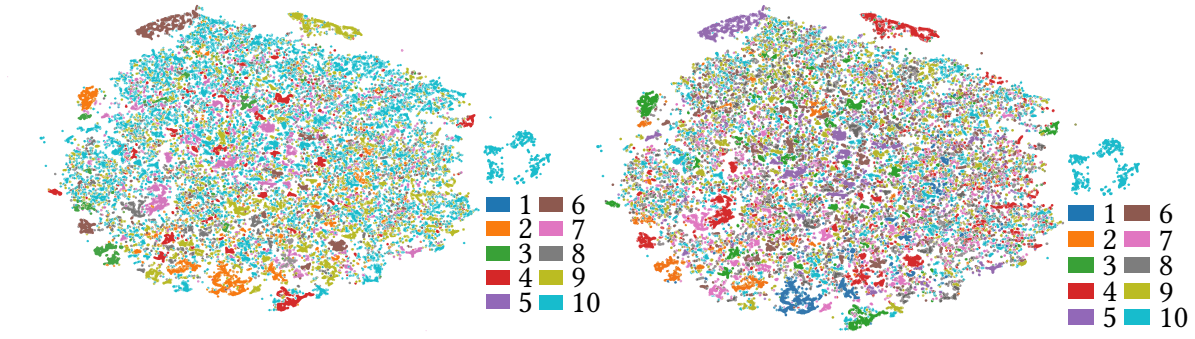


FIGURE 5.2 – Projection t-SNE en deux dimensions des fenêtres MFCC de parole sur la tâche de lecture. À gauche la coloration correspond à l'intelligibilité et à droite la coloration correspond à la sévérité (scores perceptifs de la tâche de description d'image).

Malgré cet état de fait, nous pouvons remarquer que les personnes ayant des scores élevés ont leurs fenêtres MFCC plus éparses sur ces projections. Ceci correspond à notre intuition originelle pour le score entropique de la parole que nous allons créer par la suite.

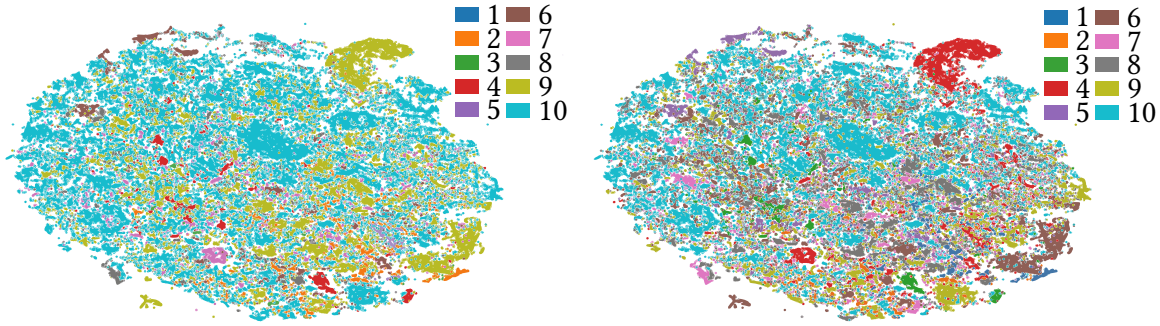


FIGURE 5.3 – Projection t-SNE en deux dimensions des fenêtres MFCC de parole sur la tâche de description d'image. À gauche la coloration correspond à l'intelligibilité et à droite la coloration correspond à la sévérité (scores perceptifs de la tâche de description d'image).

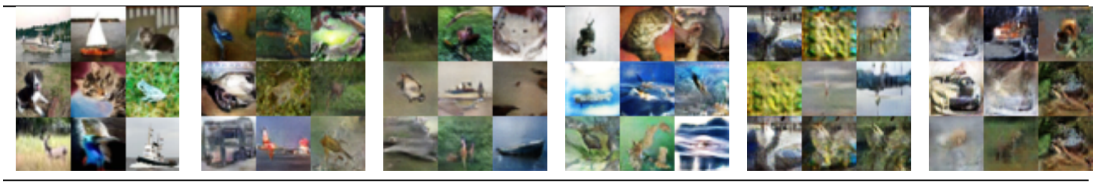
5.2 Création d'un score entropique de la parole

Dans notre méthode, notre score est une mesure entropique calculée à partir de représentations spectrales, cepstrales, ou apprises par un réseau de neurones. Ceci nous permet d'éviter d'utiliser des données de notre corpus cible et d'utiliser des corpus externes pour créer nos représentations calculées pour notre tâche.

Ainsi, nous proposons des adaptations des représentations des signaux émis par nos participants pour calculer un score entropique. Ce score est semblable à l'Inception Score (IS) utilisé pour des images [Salimans et al., 2016]. Ceci nous donne un score par participant. Ici, les scores obtenus par les contrôles servent de référence et les scores des patients doivent se rapprocher au maximum de ces scores de références. Plus le score d'un patient est éloigné de ceux de références, plus la sévérité de la maladie est considérée grande. Cette comparaison peut être vue comme une mesure de l'altération enduite par la maladie. Dans les sous-chapitres qui suivent, nous détaillerons ce score et nous montrerons et analyserons les résultats obtenus avec une telle approche.

5.2.1 Présentation de l'approche

Dans le domaine de la vision par ordinateur, la communauté scientifique utilise l'Inception Score (IS) comme une métrique permettant de voir dans quelle mesure les données générées par un modèle sont proches des données réelles [Salimans et al., 2016]. Ce score mesure la qualité des images créées par des modèles génératifs (notamment les réseaux adversaires génératifs). L'IS implique des prédictions du modèle Inception [Szegedy et al., 2016] (d'où le nom IS de l'approche qui peut porter à confusion) et est une combinaison de deux attentes entropiques sur les probabilités sorties du modèle Inception. Il permet d'évaluer les différences entre plusieurs générateurs d'images pour voir dans quelle mesure leurs performances diffèrent du score obtenu par des échantillons réels. Un exemple d'utilisation de cette mesure en image se trouve en figure 5.4.



Données réelles	Modèle 1	Modèle 2	Modèle 3	Modèle 4	Modèle 5
$11.24 \pm .12$	$7.54 \pm .07$	$6.86 \pm .06$	$6.83 \pm .06$	$4.36 \pm .04$	$3.87 \pm .03$

FIGURE 5.4 – Exemple d'utilisation de l'IS pour distinguer des images venant de cifar10 et provenant de cinq autres modèles. Figure reprise du papier original de l'IS [Salimans et al., 2016]

Similairement, nous adoptons cette approche aux locuteurs de notre corpus cancer. Nous utilisons les signaux audio pour inférer un score comparable à l'indice de sévérité des troubles de la parole. Dans notre méthode, nous faisons les hypothèses suivantes :

- chaque participant est un générateur,
- chaque participant prononce le même contenu $\mathcal{S}$ (un ensemble de phrases dans notre cas), afin de garantir que les sons produits appartiennent au même domaine et sont par conséquent comparables,
- le groupe des contrôles représente la qualité de la parole à atteindre par les patients.

Compte tenu de ces hypothèses, nous avons calculé un score pour chaque participant. La figure 5.5 illustre le traitement que nous proposons. L'entrée correspond à une séquence d'échantillons $\boldsymbol{x}$ représentant le signal produit par un locuteur l. Le locuteur pouvant produire des silences longs (par pénibilité pour prendre une pause et/ou des silences avant et après l'enregistrement du locuteur), nous utilisons un détecteur d'activité vocale (VAD dont ses implications seront discutées dans la section 5.2.2). Ce VAD extrait la séquence $\boldsymbol{x}'$ qui correspond à notre ensemble de phrases prédéfinies $\mathcal{S}$.

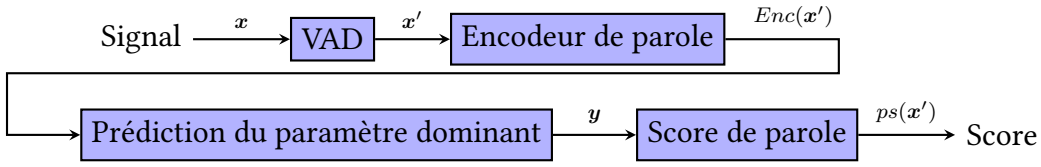


FIGURE 5.5 – Suite de traitements de notre approche non supervisée.

Les scores obtenus du groupe de contrôle représentent les scores à atteindre (comme les données réelles pour le Score d'Inception) par les patients (originellement les données factices dans l'IS, ou ici les voix dégradées). Plus le score d'un patient est proche des scores de contrôle, moins sa parole a été dégradée. L'objectif est d'obtenir une mesure de la qualité de la parole comparable à l'indice de sévérité des troubles de la parole. En comparaison à l'IS, nous avons fait quelques adaptations : nous utilisons des représentations (apprises par un modèle ou des paramètres acoustiques) au lieu d'un modèle de classification. Ceci nécessite d'adapter les sorties de ces encodeurs pour utiliser un score semblable à l'IS. Ainsi, nous normalisons et

transformons ces caractéristiques en probabilités pour mettre en avant le paramètre prédominant : plus de détails sont disponibles dans la section 5.2.4. Enfin, notre approche utilise ces probabilités (telle une sortie du modèle Inception) pour calculer un score entropique comparable à l'IS. Nous détaillerons cette partie dans la section 5.2.5.

Dans nos expériences, nous avons utilisé l'encodeur PASE+ [Ravanelli et al., 2020] comme représentation, mais nous avons également essayé des paramètres acoustiques, tels les MFCC et les MEL spectrogrammes. Néanmoins, d'autres encodeurs ou représentations du signal peuvent être utilisés, tant les caractéristiques véhiculent le même type d'information (voir la section 5.2.3 pour une description détaillée).

5.2.2 Détection d'activité vocale

La première étape de notre traitement consiste à supprimer les longs silences à l'aide d'un détecteur d'activité vocale. Soit $\boldsymbol{x}'$ les échantillons actifs d'un signal de parole extrait de VAD($\boldsymbol{x}$), VAD étant toute technique qui préserve les silences courts et la voix du locuteur (parole). Pour nos expériences, nous avons utilisé le même détecteur que celui présenté dans la section 3.3.2.

Notre méthode n'intègre pas les longues latences entre les mots (pouvant être dues à une gêne de la maladie). En effet, pour chaque locuteur, nous mesurons uniquement un score global de la qualité de la production. Même si les longs silences sont une information cruciale pour quantifier l'intelligibilité ou la sévérité, nous les évitons : il s'agit d'une conséquence de notre hypothèse que les locuteurs sont considérés comme des générateurs. Effectivement, avec notre approche, les longs silences représentent une variété de production du générateur qui peut fausser les résultats souhaités. Néanmoins, pour éviter les perturbations de l'encodeur sélectionné (lorsque nous utilisons un réseau de neurones) et inclure les silences de certains phonèmes (occlusives, plosives, etc.), nous gardons les courts silences.

Après avoir extrait le signal actif (parole), nous calculons ses représentations.

5.2.3 Encodeur de parole

L'encodeur de parole doit décrire le signal $\boldsymbol{x}'$ en une séquence de représentations aux dimensions fixes. Il doit idéalement éviter de coder les bruits environnementaux présents dans le signal et doit se concentrer sur les représentations de la parole (notre normalisation permet d'éviter les perturbations par des bruits constants).

Désignons par Enc l'encodeur de parole ou tout paramètre acoustique. Les représentations produites par Enc doivent réduire la dimensionnalité du signal d'entrée. Pour la partie expérimentale, nous avons choisi les MFCC, les Mel spectrogrammes et l'encodeur PASE+ [Ravanelli et al., 2020]. Pour les MFCC et Mel spectrogrammes, nous avons utilisé l'implémentation de librosa [McFee et al., 2015].

Nous avons choisi PASE+ [Ravanelli et al., 2020], car il est conçu pour être un encodeur de signaux vocaux générique, adapté à n'importe quelle tâche vocale, avec des trames proches du niveau phonémique. Les auteurs ont entraîné un encodeur à l'aide de plusieurs tâches autosupervisées (telles que la reconstruction du signal, la reconstruction de MFCC, la reconstruction de l'énergie, etc.) et de tâches binaires (Local Info Max et Global Info Max). Le fait de partager le même encodeur pour toutes les tâches oblige cet encodeur à représenter de

multiples paramètres de la parole. De plus, les auteurs de PASE+ ont entraîné leur modèle avec des mécanismes de débruitage qui répondent à nos besoins.

Nous détaillerons les choix de paramètres vocaux et d'encodeur dans la section 5.3.

Après avoir extrait ces encodages, nous allons transformer ces représentations pour avoir des propriétés semblables au modèle Inception.

5.2.4 Prédiction du paramètre dominant

Pour garantir un type de sortie semblable à celui du modèle Inception, utilisé pour la métrique IS, nous proposons d'adapter la sortie de l'encodeur en un classifieur de paramètre dominant. Pour calculer cette séquence de probabilités $\boldsymbol{y}$, nous proposons :

$$\boldsymbol{y} = \text{softmax}(\text{abs}(f_{\text{norm}}(Enc(\boldsymbol{x}'))))) \tag{5.1}$$

avec f_{norm} étant une fonction de normalisation établie sur les espaces latents correspondant au signal vocalisé ($Enc(\boldsymbol{x}')$).

Dans nos expériences, nous avons testé plusieurs normalisations : L_1, $L_{\max}$, L_∞. Nous avons appliqué ces normalisations pour chaque dimension de nos différents pas de temps. Ainsi pour une dimension $\boldsymbol{z} = [\boldsymbol{z}_1, ..\boldsymbol{z}_n]$ nous avons :

$$||\boldsymbol{z}||_1 = \boldsymbol{z} / \sum_{i=1}^{n} |z_i| \tag{5.2}$$

$$||\boldsymbol{z}||_{\max} = \boldsymbol{z} / \max_i z_i \tag{5.3}$$

$$||\boldsymbol{z}||_\infty = \boldsymbol{z} / \max_i |z_i| \tag{5.4}$$

Nous avons également essayé le $zscore$. Pour le cas du $zscore$, nous calculons la standardisation comme suit :

$$zscore(\boldsymbol{X}'_t) = \frac{\boldsymbol{X}'_t - \mu_{\text{fichier}}}{\sigma^2_{\text{fichier}}} \tag{5.5}$$

où μ_{fichier} et σ_{fichier} sont la moyenne et l'écart-type des espaces latents (données par Enc) du fichier audio de la lecture d'un participant. Ici $\boldsymbol{X}'_t$ désigne la représentation latente d'un pas de temps t ($Enc(\boldsymbol{x}_t)$).

L'utilisation de la fonction absolue dans l'équation 5.1 permet aux valeurs fortement négatives d'avoir un impact équivalent aux valeurs positives sur le score entropique. Effectivement, toutes les deux portent des informations significatives du signal. Cet élément est important dans le cas d'utilisation d'un réseau de neurones comme encodeur. Nous utilisons ensuite la fonction softmax pour mettre en évidence le paramètre le plus présent dans chaque fenêtre temporelle selon l'encodage d'Enc. Dans nos expériences, Enc représente soit des paramètres acoustiques (tels que les MFCC ou les Mel spectrogrammes) soit un encodeur de caractéristiques (ici PASE+). Ensuite, nous calculons notre score entropique de parole, similairement à l'IS.

5.2.5 Score entropique de parole

Soit y le vecteur de probabilité correspondant à l'espérance des probabilités d'$Enc(\boldsymbol{x}')$ et y_t le vecteur de probabilité pour un pas de temps ($Enc(\boldsymbol{x}'_t)$). Pour calculer notre score de production ps (inspiré par l'IS), nous procédons comme suit :

$$ps(\boldsymbol{x}') = \exp(\rho_{\boldsymbol{x}'_t}(\mathbf{KL}(p(y_t|\boldsymbol{x}'_t)||p(y)))) \tag{5.6}$$

avec $\rho_{\boldsymbol{x}'_t}$ étant une fonction de statistique descriptive et **KL** étant la divergence de Kullback-Leibler.

Pour nos expériences, nous nous sommes limités aux fonctions suivantes pour $\rho_{\boldsymbol{x}'_t}$: espérance $\mathbb{E}$, écart type σ et médiane. Néanmoins, d'autres fonctions statistiques sont possibles. Notez que si $\rho_{\boldsymbol{x}'_t} = \mathbb{E}_{\boldsymbol{x}'_t}$, nous avons la formulation originale de l'IS [Salimans et al., 2016]. Après avoir fixé $\rho_{\boldsymbol{x}'_t}$, nous pouvons calculer le score entropique de parole pour chaque locuteur.

Nous avons choisi cette formulation, car elle implique que les productions d'un texte doivent simultanément être : localement stables (idéalement proches d'unités phonétiques) et globalement variées (les participants doivent être capables de prononcer différentes unités sonores). Ainsi, la représentation produite par l'encodeur doit être clairsemée (représentation « sparse »).

5.3 Résultats du score entropique de la parole

Pour nos expériences, nous avons utilisé le corpus cancer et la tâche de lecture pour garantir que chaque locuteur prononce le même message. De plus, nous avons essayé plusieurs encodages du signal de la parole dont : les MFCC, les Mel spectrogrammes et la sortie de l'encodeur PASE+.

Pour notre expérience établie sur PASE+, nous avons choisi d'utiliser le modèle appris par les auteurs sur 50 h d'anglais du corpus Librispeech [Panayotov et al., 2015]. Ainsi, lors de l'utilisation de ce modèle, notre approche suppose que :

1. l'indice de sévérité est un concept universel : il devrait donc y avoir des similitudes considérables entre l'anglais et le français,
2. le modèle encode une représentation suffisamment faible (proche du signal) de la parole pour être utile aux données françaises.

Un résumé des meilleurs scores obtenus par encodeur est disponible en table 5.1. Ainsi, notons que le score utilisant le modèle PASE+ nous donne de meilleurs résultats. Les Mel spectrogrammes nous permettent d'obtenir de meilleurs résultats que les MFCC (en comparaison aux approches de « few-shot » du chapitre précédent). Grâce à l'encodage PASE+, notre méthode est plus performante que celle de [Balaguer et al., 2019] et ne nécessite pas l'utilisation d'un aligneur forcé. Cette meilleure performance peut être due au débruitage de l'encodeur PASE+ : celui-ci encode peut-être plus d'informations extra-acoustiques que les MFCC et les Mel spectrogrammes. Cela est d'autant plus remarquable, que le modèle PASE+ a été appris sur de l'anglais. Ainsi, ce genre de modèles encode des caractéristiques de suffisamment bas niveau et génériques pour être appliquées sur d'autres langues.

TABLE 5.1 – Résumé des résultats obtenus pour chaque encodeur.

	MFCC	Mel spectrogrammes	PASE+
Corrélation sur les participants	-0,697	0,800	-0,868
Corrélation uniquement sur les patients	-0,637	0,724	-0,829

Les résultats détaillés, compte tenu du ρ et du f_{norm} choisis, sont disponibles dans la table 5.2 pour les paramètres MFCC, dans la table 5.3 pour les Mel spectrogrammes et dans la table 5.4 pour l'encodeur PASE+.

Le premier élément ressortant de nos résultats est que nous ne pouvons pas facilement recommander un ρ ou un f_{norm} fixes. En effet, les résultats sont très variables selon l'encodeur choisi.

Il est à noter que lorsque la sortie de l'encodeur est entièrement positive, $f_{\text{norm}} = l_{max}$ est équivalent à l_∞. Ceci est le cas pour nos résultats avec les Mel spectrogrammes comme encodeur.

TABLE 5.2 – Détail des résultats obtenus avec les MFCC comme encodeur. Avec à gauche les corrélations sur tous les participants et à droite les corrélations pour les patients uniquement.

$\rho \setminus f_{\text{norm}}$	l_1	l_{max}	l_∞	$zscore$
σ	0,672		0,444	0,311
$\mathbb{E}$	0,674		0,347	0,313
médiane	-0,697	-0,102	-0,436	-0,281

$\rho \setminus f_{\text{norm}}$	l_1	l_{max}	l_∞	$zscore$
σ	0,573		0,328	0,405
$\mathbb{E}$	0,570		0,214	0,398
médiane	-0,637	-0,092	-0,352	-0,350

TABLE 5.3 – Détail des résultats obtenus avec les Mel spectrogrammes comme encodeur. Avec à gauche les corrélations sur tous les participants et à droite les corrélations pour les patients uniquement.

$\rho \setminus f_{\text{norm}}$	l_1	l_{max} / l_∞	$zscore$
σ	0,798	0,433	0,017
$\mathbb{E}$	0,800	0,238	0,043
médiane	-0,500	0,183	0,278

$\rho \setminus f_{\text{norm}}$	l_1	l_{max} / l_∞	$zscore$
σ	0,722	0,303	0,004
$\mathbb{E}$	0,724	0,140	-0,020
médiane	-0,393	0,220	0,278

TABLE 5.4 – Détail des résultats obtenus avec PASE+ comme encodeur. Avec à gauche les corrélations sur tous les participants et à droite les corrélations pour les patients uniquement.

$\rho \setminus f_{\text{norm}}$	l_1	l_{max}	l_∞	$zscore$
σ	0,648	0,699	-0,404	-0,234
$\mathbb{E}$	0,384	0,798	-0,414	-0,868
médiane	-0,448	0,823	0,682	0,802

$\rho \setminus f_{\text{norm}}$	l_1	l_{max}	l_∞	$zscore$
σ	0,531	0,581	-0,508	-0,206
$\mathbb{E}$	0,236	0,685	-0,511	-0,829
médiane	-0,350	0,770	0,680	0,771

Les nuages de points pour les meilleurs résultats de chaque encodeur sont disponibles dans les figures 5.6, 5.7 et 5.8. Dans ces figures, chaque point représente un fichier où un participant lit les mêmes phrases que les autres participants.

Nous remarquons que le nuage de points fondé sur PASE+ est plus diffus pour les patients avec une faible sévérité que pour nos résultats fondés sur les Mel spectrogrammes. Néanmoins, pour les scores avec une sévérité plus haute, le score utilisant PASE+ est le moins diffus. Ainsi, une fusion du score fondé sur les Mel spectrogrammes et du score fondé sur PASE+ pourrait améliorer nos résultats.

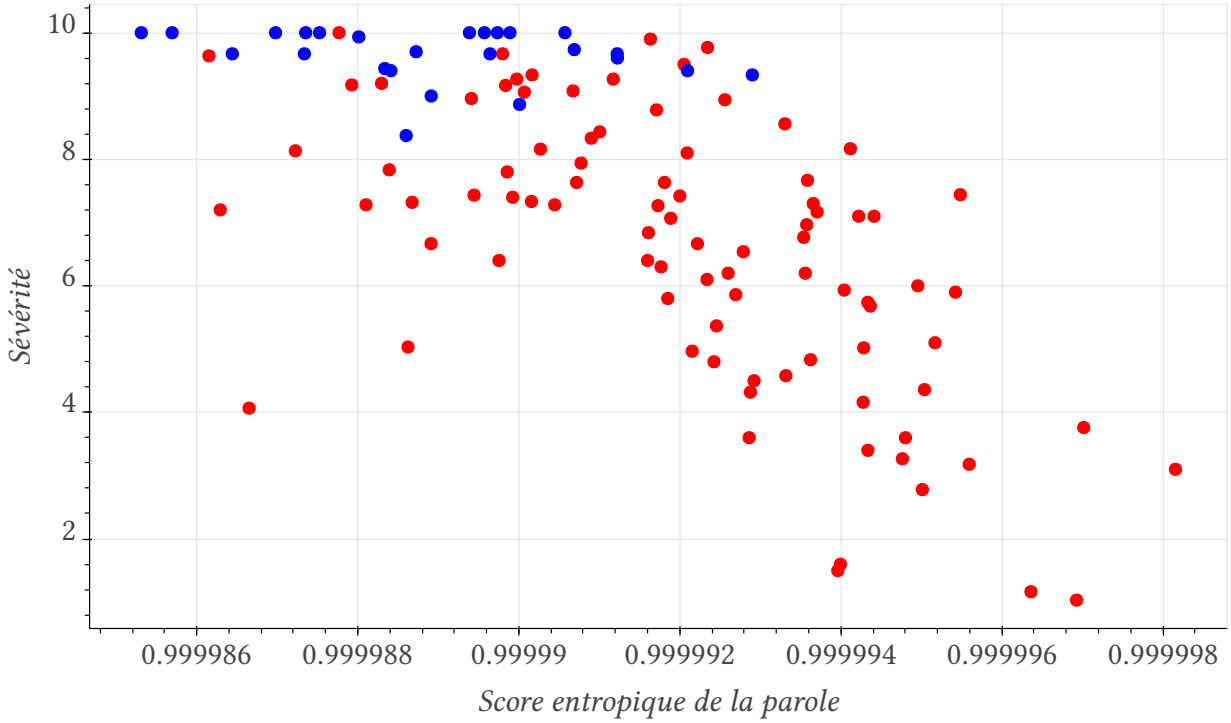


FIGURE 5.6 – Nuage de points du meilleur score entropique de la parole s'appuyant sur les MFCC. Les points rouges correspondent aux fichiers patients, tandis que les bleues correspondent aux fichiers contrôles.

Notez que vous trouverez sur GitHub des résultats plus détaillés des expériences, y compris les nuages de points interactifs [28] ainsi que le code utilisé [29].

Au sein du projet RUGBI deux autres approches ont été développées sur la tâche de lecture par [Quintas et al., 2020] et [Abderrazek et al., 2020]. La première est semblable à l'approche de Laarid et al. [Laaridh et al., 2017] (vu en chapitre 2). Néanmoins, Quintas et ses collègues utilisent les x-vecteurs à la place des i-vecteurs. Ces derniers sont fondés sur le même principe que les i-vecteurs, mais utilisent l'apprentissage profond pour modéliser les caractéristiques du locuteur. Ainsi, les auteurs ont obtenu une corrélation de 0,85 avec l'intelligibilité. Bien que l'indice de sévérité mesure des altérations non décelables par l'intelligibilité [Auzou et al., 2007] la prédiction d'intelligibilité peut être un atout pour évaluer l'indice de sévérité. En regardant plus en détail leur résultat, on s'aperçoit que leur approche a des difficultés à différencier les scores inférieurs à cinq de ceux compris entre cinq et huit.

28. https://github.com/vroger11/SAMI/tree/master/c2si_results
29. https://github.com/vroger11/SAMI

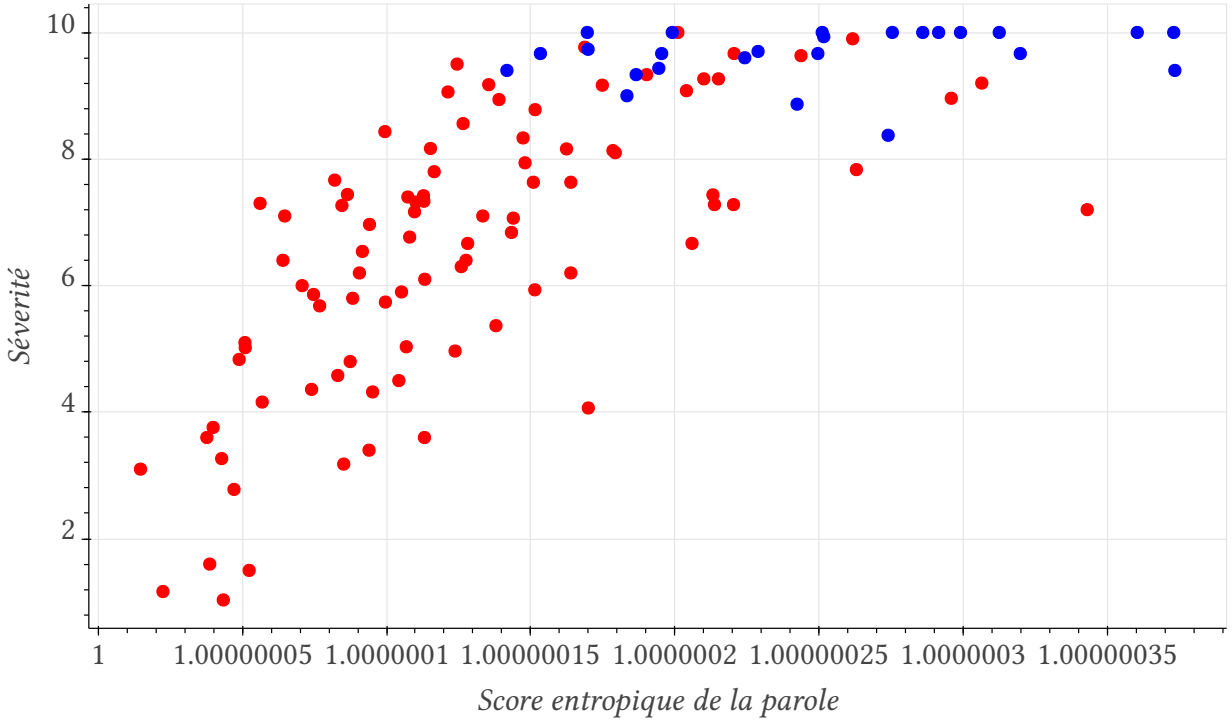


FIGURE 5.7 – Nuage de points du meilleur score entropique de la parole s'appuyant sur les Mel spectrogrammes. Les points rouges correspondent aux fichiers patients, tandis que les bleues correspondent aux fichiers contrôles.

Abderrazek et al. utilisent une approche semblable aux méthodes fondées sur les vraisemblances de phonèmes prédits par un modèle automatique [Middag et al., 2008, Maier et al., 2009]. Ici, cependant, les auteurs utilisent un modèle de réseaux profonds pour réaliser la reconnaissance de phonèmes. Ce modèle convolutif a été appris sur le corpus BREF [Lamel et al., 1991] (un corpus français contenant très peu de bruits environnementaux). En faisant ainsi, ils surpassent les résultats initiaux obtenus sur le corpus cancer [Balaguer et al., 2019]. Ainsi, à l'aide d'un alignement forcé (le même qu'utilisé dans nos expériences de few-shot), les auteurs ont calculé une précision équilibrée de reconnaissance de phonème pour obtenir une corrélation de 0,91. Cette corrélation surpasse celle de notre approche. En observant plus en détail leurs résultats, leur approche possède des difficultés à différencier les scores inférieurs à cinq. Bien que notre approche n'obtienne pas une aussi bonne corrélation globale, la diffusion des faibles scores est similaire (voir plus diffuse que les hauts scores) lorsque nous utilisons l'encodeur PASE+. Notre approche ne nécessite pas l'utilisation d'un alignement forcé malgré le fait qu'il est nécessaire que les patients prononcent (ici lisent) les mêmes phrases. De plus pour la création du score de leur approche, il est nécessaire de connaître les mots que les participants sont sensés prononcer pour calculer le taux d'erreur de chaque participant. Ce pré-requis n'est également pas nécessaire avec notre approche ainsi que celle de Quintas et ses collègues. Enfin, notre système est potentiellement plus robuste aux bruits environnementaux grâce à l'utilisation d'un encodeur appris avec des techniques de débruitage. Ce dernier point

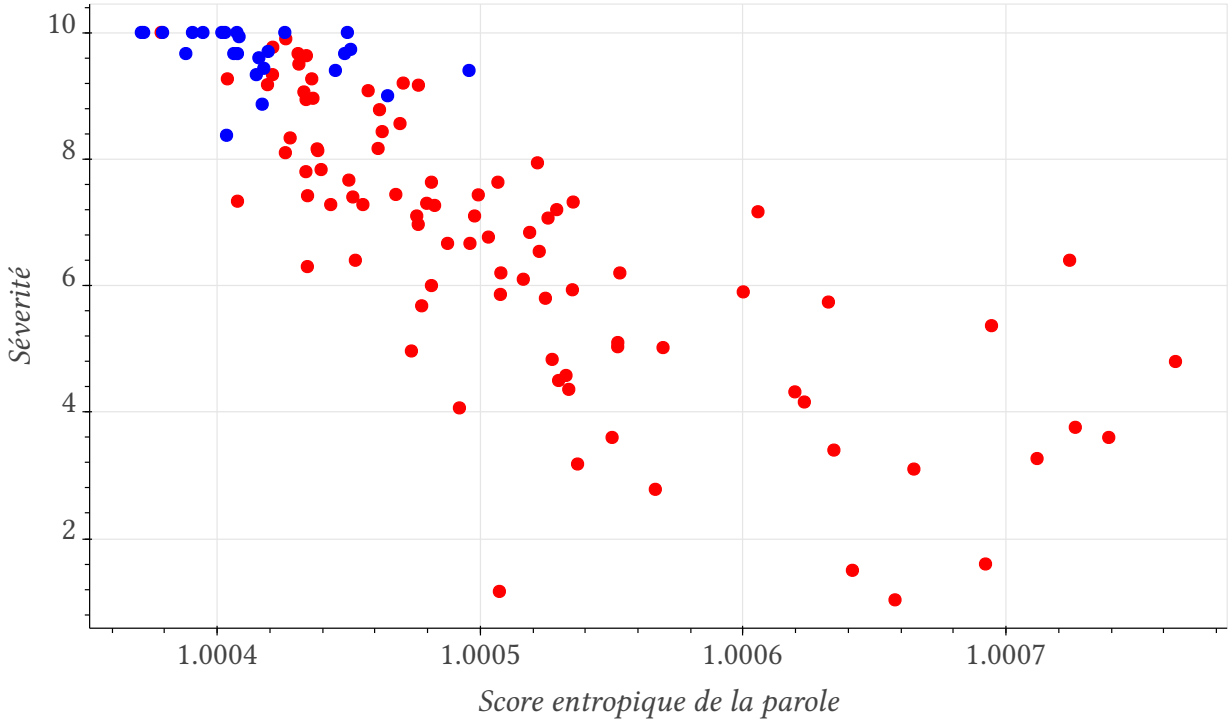


FIGURE 5.8 – Nuage de points du meilleur score entropique de la parole s'appuyant sur PASE+. Les points rouges correspondent aux fichiers patients, tandis que les bleues correspondent aux fichiers contrôles.

est un avantage pour imaginer notre approche au sein de l'hôpital dont l'environnement n'a pas toujours un rapport signal sur bruit favorable comme pour notre corpus cancer.

5.4 Discussion de nos résultats

Dans ce chapitre, nous avons proposé une nouvelle méthode pour évaluer une mesure de la parole semblable à l'indice de sévérité des troubles de la parole chez les patients. Notre méthode consiste à adapter de la sortie d'encodeurs de représentations du signal de parole pour réaliser un score entropique de la parole afin de caractériser la dégradation de la parole des patients. Dans notre meilleure solution, nous utilisons un encodeur pré-entraîné sur de la lecture de texte anglais (PASE+) pour construire une métrique qui évalue la qualité de la production vocale. Ce type de modèle nous garantit une certaine robustesse aux bruits environnementaux, aux locuteurs. Bien que notre approche reprenne l'idée de prendre les contrôles comme référence, comme dans Janbakhshi et al. [Janbakhshi et al., 2019] (vu dans le chapitre 2), nous proposons une approche pouvant utiliser des modèles qui permettent d'avoir une robustesse aux bruits, aux accents et intra-locateur. Nous avons utilisé notre méthode sur un corpus français malgré l'utilisation d'un encodeur anglais (non adapté/affiné sur de la parole française). Ceci nous garantit une certaine robustesse aux accents locaux. L'acquisition de nouvelles données médicale étant difficile (par pénibilité et fatigabilité pour les patients), nous

n'avons pas pu affiner les paramètres du modèle sur de la parole pathologique. Néanmoins, nous obtenons une corrélation de Spearman de 0,87, ce qui encourage l'utilisation d'une telle méthode pour les applications médicales du monde réel. Ceci indique que la mesure de la sévérité des troubles de la parole peut être semblable pour le français et pour l'anglais. Ainsi, nous montrons que certains encodeurs de réseaux neuronaux profonds peuvent être utilisés sans modification dans des tâches ne disposant pas de suffisamment de données pour réaliser un apprentissage ou un affinage des paramètres du modèle. Ces résultats nous ont encouragés à créer une application médicale pour mettre cette approche en situation réelle.

6

Application clinique

Durant ma thèse, j'ai décroché un financement lors de la campagne de pré-maturation Protopitch 2020 lancée par la Société d'accélération du transfert de technologies Toulouse Tech Transfert (SATT TTT). Ce financement a été utilisé pour valoriser mes travaux de thèse en recrutant un ingénieur d'études [30]. J'ai encadré ce travail visant à réaliser une application mobile iOS utilisable dans le milieu hospitalier.

Suite à un recueil des besoins, les points suivants étaient les éléments principaux requis par les aides soignants :

— le score produit doit être semblable à l'indice de sévérité de la parole,
— l'application se doit d'être simple d'utilisation,
— l'application doit permettre le suivi de patients.

Ce travail a permis de réaliser une application de suivi des cancers ORL. L'objectif de celle-ci est de mesurer l'évolution de l'indice de sévérité des patients. De ce fait, nous pourrons disposer d'un indicateur calculé rapidement après un enregistrement de la parole et voir l'effet des chimiothérapies, de la rééducation et/ou des traitements médicamenteux sur la production de parole des patients. Pour répondre aux différents besoins, nous avons choisi de développer l'application pour tablette [31].

6.1 Fonctionnalités

Cet outil permet au praticien de mesurer l'ensemble de ces patients (au travers de l'extrait de la chèvre de monsieur Seguin), de les gérer via l'application et d'afficher l'historique de ces derniers. La figure 6.1 regroupe un exemple de chacune de ces fonctionnalités.

Pour mesurer un patient, le praticien doit sélectionner le profil de ce dernier (dont l'identifiant correspond à celui utilisé au sein du CHU [32]), le patient prend ensuite la tablette et clique sur une icône d'enregistrement pour lire l'extrait de la chèvre de monsieur Seguin (il obtient

30. CDD à temps partiel réalisé par Gauthier Arcin.
31. L'hôpital prévoyant d'utiliser ce support prochainement, ce choix nous a semblé être le plus pertinent.
32. Ceci aide les praticiens à retrouver plus facilement leurs patients comme ils utilisent majoritairement cet identifiant pour les retrouver au travail de tous les outils numériques.

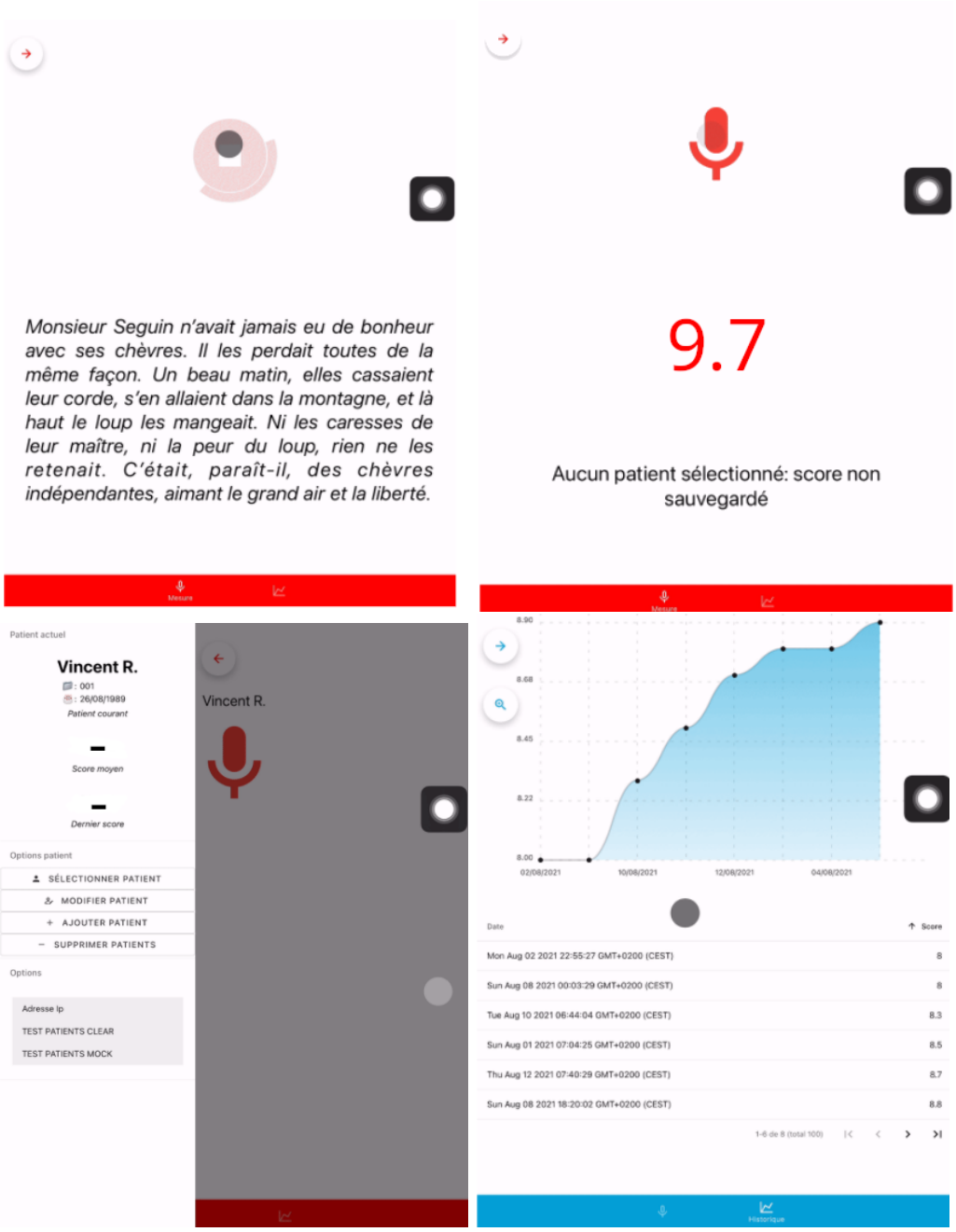


FIGURE 6.1 – Exemple factice (aucun patient n'est exposé sur ces figures) de l'interface de l'application avec de gauche à droite et de haut en bas, la lecture de l'extrait de la chèvre de monsieur Seguin, l'affichage d'un score, la gestion de patients et l'affichage de l'historique d'un patient.

l'écran qui se trouve en haut à gauche de la figure 6.1). Un score de sévérité est obtenu [33]

33. Ce score utilise notre meilleure approche vu en chapitre 5. Une régression linéaire établie sur une approximation Nystroem de 5 composantes nous permet de transformer le score obtenu en un score de 0 à 10 (avec l'utilisation de seuils pour éviter de dépasser ces bornes).

(voir l'écran situé en haut à droite de la figure 6.1). Pour éviter les silences liés à la latence des utilisateurs utilisant l'interface (le patient doit aussi appuyer sur l'icône stop une fois l'enregistrement fini), nous utilisons notre détecteur d'activité vu en chapitre 3 pour supprimer les silences de début et de fin des enregistrements.

Après avoir réalisé plusieurs mesures de ses patients, le praticien peut visualiser l'historique de chaque patient (avec filtrage par date de possible, voir en bas à droite de la figure 6.1) pour observer l'évolution de la production de parole pour chaque patient.

Enfin, nous avons réalisé une dernière caractéristique de notre outil est l'évolutivité. Ceci a été assuré en proposant une interface de programmation simple pour qu'un chercheur puisse simplement ajouter une mesure. En effet, l'ajout d'un nouveau module (suivant un prototype donné à l'avance dans la documentation de notre outil) suffit pour obtenir une nouvelle mesure sur l'application. L'utilisateur de la tablette n'a plus qu'à sélectionner la mesure qu'il veut afficher dans l'application. Ceci pourra permettre de comparer les différentes mesures développées sur un environnement réel. Cette fonctionnalité facilitera la collaboration entre les équipes de recherche, en identifiant les mesures les plus appropriées pour le suivi des patients subissant des troubles de la parole.

6.2 Réalisation

Pour matérialiser cette application, nous avons choisi un fonctionnement avec un client (la tablette) et un serveur (machine située à l'hôpital).

Le serveur réalise la gestion des patients et le calcul des scores (dont les communications avec les tablettes sont cryptées en AES 256 [34]). Ce dernier est codé en Python à l'aide de la librairie Flask [35] pour la gestion des données. Il est déployé à l'aide d'un conteneur docker [36] pour faciliter sa mise en production et sa maintenance. L'avantage d'utiliser un serveur est que nous pouvons utiliser un GPU. Ainsi, nous pouvons déployer plus facilement et sans adaptation des modèles contenant des codes écrits en CUDA [37] et avoir des retours pour voir si l'adaptation du modèle pour smartphone et/ou tablette est intéressante ou non.

L'application mobile (servant de rôle de client) a été développée pour des iPad mini. Ce choix de tablette a été motivé par la qualité et l'homogénéité des traitements du microphone du dispositif dans l'optique d'une utilisation à grande échelle de notre solution. Pour ce faire, l'application est développée en TypeScript [38] et utilise le framework React Native [39]. De ce fait, l'application peut fonctionner sous Android, mais nos efforts se sont concentrés sous iOS.

Deux documentations ont été réalisées en relation avec notre application. Une première destinée aux praticiens (l'utilisateur final) et une seconde pour le développement de l'application et du serveur (organisation de l'architecture, installation...). De plus, des tests unitaires et

34. Pour plus de détails sur ce protocole, vous pouvez vous référer au document suivant : https://csrc.nist.gov/csrc/media/publications/fips/197/final/documents/fips-197.pdf
35. https://flask.palletsprojects.com/
36. https://www.docker.com/
37. https://developer.nvidia.com/cuda-toolkit
38. https://www.typescriptlang.org/
39. https://reactnative.dev/

d'intégration automatiques ont été établis pour assurer une certaine fiabilité dans le développement. La présentation utilisée pour introduction cette application au CHU est disponible en annexe A.

Après de nombreux problèmes techniques, l'application est maintenant déployée au sein du CHU et est utilisable. Elle va donner lieu à des expérimentations pour en évaluer la viabilité dans un contexte clinique. Nous attendons maintenant les premiers retours pour appréhender les performances en conditions réelles de notre application.

Conclusion de partie

Nous avons dans un premier temps réalisé une revue des techniques supervisées et du lien entre quantité de données, performances et complexité du modèle. Suite à cela, nous avons utilisé des techniques supervisées dites de few-shot pour les adapter à la détection de phonèmes. Ceci a engendré à la publication d'un papier dans le journal Eurasip [Roger et al., 2022a]. Nos premières expériences pour faire apprendre le concept de sévérité de la parole avec ces techniques nous ont donné deux indications :

- la sévérité est plus facile à apprendre sur les voyelles. Ainsi, les membres du jury de notre corpus se sont peut-être focalisés sur la prononciation des voyelles pour réaliser leurs évaluations. Une étude avec plus de données et établi sur d'autres textes est nécessaire pour affirmer que ce biais est récurant dans l'évaluation d'experts en clinique.
- l'apprentissage avec seulement des voix d'hommes donne des résultats similaires à un apprentissage complet (utilisation de voix d'hommes et de femmes). Ceci nous fait penser que les caractéristiques du signal liées à l'indice de sévérité sont semblables pour les hommes et femmes, et peut être appris indépendamment du sexe.

Le manque de données pour des approches supervisées (ensembles d'entraînements et d'évaluation trop faibles) nous a emmenés vers d'autres approches. Nous avons ainsi développé une mesure entropique de la parole se fondant sur un modèle auto-supervisé (ici PASE+). Les avantages par rapport aux autres méthodes de l'état de l'art sont les suivants :

- il n'est pas nécessaire pour le modèle de connaître les mots des phrases prononcées (bien que chaque utilisateur doit prononcer la même phrase pour pouvoir les comparer),
- aucun alignement forcé n'est utilisé,
- robustesse de nos modèles aux bruits environnementaux, aux variations intra-locuteurs dans le temps et aux accents.

Cette approche a été publiée et présentée aux Journées d'Études sur la Parole [Roger et al., 2022b]. Suite à ces résultats, j'ai décroché une bourse auprès de TTT pour réaliser une application tablette utilisant cette méthode. La tablette et l'application sont maintenant installées au CHU de Toulouse et nous attendons les retours des cliniciens.

Conclusion et perspectives

Conclusion

Durant cette thèse, j'ai cherché à modéliser le concept de sévérité de la parole pathologique à l'aide de techniques d'apprentissage moderne (principalement fondé sur des réseaux de neurones). Pour mon travail de thèse, je devais répondre à la problématique suivante : *est-ce que les techniques d'apprentissage peuvent, avec un corpus de données limité, modéliser le concept d'indice de sévérité du trouble de la parole tout en étant robustes ? Et, si oui, serait-il possible d'envisager une application médicale à ces travaux ?*

Dans un premier temps, j'ai analysé le corpus cancer que nous avions à disposition. J'ai ainsi contribué à un article de journal [Woisard et al., 2021]. Mon travail a permis de vérifier les équilibres de répartition des différents locuteurs et a mis en exergue l'importance du paramètre de la durée de lecture : celui-ci obtient de bonnes corrélations avec les scores perceptifs de sévérité et d'intelligibilité. De plus, cette tâche garantit une assez bonne couverture de l'inventaire phonétique français, et nous assure de disposer d'une normalisation de la quantité de données à traiter pour chaque participant. Nous nous sommes donc focalisés sur la tâche de lecture pour la suite de mes travaux.

Après avoir réalisé un état de l'art des systèmes récents de traitement de la parole (reconnaissance de parole, reconnaissance des émotions et identification du locuteur), nous avons essayé de modéliser le concept de sévérité directement au niveau acoustique. Pour cela, une segmentation en phonèmes a été utilisée, pour avoir une analyse fine du concept. Nos premiers essais se sont révélés infructueux, aussi bien avec des apprentissages de réseaux de neurones profonds qu'avec l'utilisation d'apprentissage par transfert. En effet, les systèmes de l'état de l'art requièrent une grande quantité de données (de l'ordre du millier d'heures) et ne sont pas adaptés à notre situation.

Suite à ce constat, nous nous sommes tournés sur l'apprentissage dit « few-shot » (basé sur l'utilisation de très peu d'exemples représentatifs). Ce choix a été motivé par de récents travaux qui ont montré un intérêt de ces techniques pour les enregistrements audio [Eloff et al., 2019, Feng and Chaspari, 2021, Anand et al., 2019, Zhang et al., 2019]. Afin de nous comparer à l'état de l'art, et valider le choix de cette technique, nous avons sélectionné le corpus TIMIT pour réaliser des expérimentations. Nos premiers résultats indiquent que le « few-shot » est possible pour la reconnaissance de phonèmes. En effet, nous avons réussi à obtenir une précision de 41% en utilisant seulement 1% des données disponibles pour l'apprentissage du corpus TIMIT. De plus, ce résultat a été obtenu sans augmentation de données. Or, les méthodes de reconnaissance avec lesquelles nous nous comparons utilisent ce mécanisme pour obtenir les

meilleurs résultats. Ce travail, combiné à l'état de l'art des systèmes récents de traitement de la parole et d'approches few-shot, a donné lieu à l'acceptation d'un article dans EURASIP Journal on Audio, Speech, and Music Processing [Roger et al., 2022a].

Suite à ces premiers résultats, nous avons utilisé un réseau prototypique pour apprendre des classes établies sur l'index de sévérité de la parole. Ainsi, nous avons identifié que les voyelles semblent plus simples à apprendre (sur notre corpus cancer) que les autres phonèmes. Le corpus étant majoritairement constitué de voyelles, lors de l'annotation, les experts peuvent être très influencés par la prononciation de celles-ci : cela demande une étude plus approfondie pour être confirmée. Bien que ces résultats soient prometteurs, cette approche n'obtient pas de résultats suffisamment satisfaisants pour envisager une application médicale. Nous pourrions utiliser l'augmentation de données pour améliorer les résultats. Cependant, il serait alors nécessaire de réaliser une mesure d'impact : voir si le score attribué par les experts reste consistant.

Notre corpus étant limité en quantité de données, la nécessité d'utiliser une partie de ces données pour l'apprentissage d'un modèle est un handicap. Ainsi, après notre analyse au niveau phonétique, nous avons essayé une approche plus globale (au niveau des mots et des phrases) en proposant un score entropique. Cette approche nous permet, grâce à l'adaptation de la sortie d'encodeurs de parole, de calculer un score afin de caractériser la dégradation de la parole des patients. Dans notre meilleure solution, nous utilisons un encodeur pré-entraîné sur diverses tâches de parole (PASE+) pour construire une métrique évaluant la qualité de la production vocale. L'avantage d'utiliser un tel encodeur nous permet d'envisager des essais en situations réelles, comme cet encodeur possède une robustesse aux bruits environnementaux et une robustesse aux accents de locuteurs. Ainsi, nous obtenons une très forte corrélation de Spearman ($\rho = 0,87$) entre notre score et l'indice de sévérité issu des experts. L'avantage de notre approche est que nous n'avons pas besoin d'adapter notre métrique avec des données propres au corpus traité. De plus, nous pouvons utiliser l'ensemble des données de notre corpus pour faire des tests : nous utilisons des corpus externes pour créer un modèle servant à notre mesure entropique de la qualité de la parole. Ce travail a donné lieu à la publication d'un papier aux Journées d'Études sur la Parole (JEP 2022) [Roger et al., 2022b].

Enfin, la dernière contribution liée à ma thèse a été la création d'une application mobile fondée sur ma meilleure approche. Cette application permet à l'aide-soignant de gérer l'ensemble de ses patients pour mesurer (avec l'aide de mon indice de détérioration de la parole) et afficher un historique (suivi des patients). Cette application est déployée au CHU de Toulouse et nous attendons les retours des aide-soignants.

Mon travail de thèse m'a permis d'appréhender la difficulté de travailler avec peu de données. En effet, j'ai dû m'adapter aux données que j'avais à disposition, avec leurs avantages et inconvénients (le corpus étant créé au début de ma thèse). J'ai appris à recenser un maximum de solutions et choisir quelles sont les plus prometteuses. De plus, la recherche de financement pour la création de l'application mobile m'a apporté des compétences de recherche de fonds (réponse à un appel d'offres, présentation d'un projet et réalisation de ce dernier). Enfin, ma thèse m'a permis de développer des capacités d'adaptation. En effet, j'ai transposé des techniques provenant de la communauté d'image à la communauté de la parole. Citons, par exemple, le score Inception et les approches dites de « few-shot » [Roger et al., 2022b, Roger et al., 2022a].

Perspectives de recherche

Perspectives directes de mes travaux

Bien que notre approche obtienne une bonne corrélation sur notre corpus cancer, l'utilisation d'un corpus plus complet nous semble nécessaire. En effet, il manque des représentants ayant un faible score de sévérité (inférieur à cinq). Ceci permettrait de mieux discerner les forces et faiblesses des approches développées au sein du projet RUGBI et d'envisager une fusion de ces dernières. De plus, l'utilisation de textes mieux conçus pour la tâche de lecture (tel que préconisé dans [Pommée et al., 2022]) permettrait une analyse plus fine de ces résultats.

Notre approche ne nécessitant aucun affinage du modèle utilisé, il serait intéressant de tester notre approche sur d'autres corpus contenant d'autres types de pathologies. Nous pensons prioritairement à effectuer des essais sur le sous-corpus des patients atteints de la maladie de Parkinson du projet RUGBI.

PASE+ étant appris sur de l'anglais, il serait intéressant de pouvoir disposer d'une version française de ce modèle. Dans cet objectif, nous avons entraîné un modèle utilisant 50 heures de parole lue provenant du projet Common Voice de Mozilla [40] [Ardila et al., 2020]. Néanmoins, nous n'obtenons pas de meilleurs résultats que la version anglaise apprise sur le corpus librispeech (avec un point de moins de corrélation sur le corpus cancer). Une suite logique serait également de tester d'autres encodeurs de parole tels que Wav2Vec2.0 et HuBERT et de réaliser l'apprentissage sur un corpus français de ces approches.

PASE+ utilise des mécanismes de débruitages et nous permet d'obtenir de bons résultats avec notre approche de mesure entropique. Vu que nous avons obtenu de bons scores avec ce modèle, nous pourrions supposer que les augmentations utilisées dans ce modèle ne changeraient pas la perception du jury. Ainsi, la réverbération, les bruits additifs, le masquage temporel et fréquentiel, le clipping et les chevauchements de parole (augmentations utilisées dans PASE+) semblent intéressants pour les essayer avec nos modèles « few-shot » développés pendant la thèse. D'autres augmentations nous semblent intéressantes comme l'ajout de bruit environnemental, par exemple en provenance d'hôpitaux, pour assurer une certaine robustesse de notre approche. Néanmoins, les autres augmentations classiquement utilisées en traitement de la parole (tel que l'étirement temporel [Nguyen et al., 2020]) pourraient nécessiter une vérification de non-modification du score perçu par le jury d'experts (ce qui pourrait être coûteux et finalement non réalisable en pratique).

Une autre perspective consisterait à utiliser les encodeurs tels que PASE+, Wav2Vec2.0 ou HuBERT en entrée de nos approches de « few-shot ». En effet, pour notre score entropique, les MFCC et les Mel spectrogrammes n'ont pas permis d'obtenir un meilleur résultat que l'encodeur PASE+. Ceci peut indiquer que l'encodeur PASE+ a une représentation plus intéressante pour la création d'un indice de sévérité. Ainsi, nos premiers résultats utilisant les approches « few-shot » peuvent certainement bénéficier du pré-entraînement de ces encodeurs au lieu d'utiliser de l'augmentation de données (que les encodeurs pré-entraînés utilisent déjà).

Enfin, nous pouvons imaginer que notre application développée pour le CHU de Toulouse soit intégrée dans des plateformes plus larges telles que MonPaGe [Pernon et al., 2020] pour permet une évaluation perceptive et acoustique des troubles moteurs de la parole.

40. https://commonvoice.mozilla.org/fr

Perspectives générales

L'utilisation de mesures de dégradations de la parole permet d'envisager la sélection de modèles adaptés pour plusieurs niveaux de dégradations de la parole. En effet, la reconnaissance d'émotions ou de mots prononcés par des personnes dont la parole est altérée par la maladie est une tâche difficile, car nous manquons de données pour ces maladies [Latif et al., 2020]. Nous pouvons imaginer qu'une mesure d'altération de la parole aiderait à avoir une stratégie pour sélectionner le modèle le plus adapté. Ainsi, une telle mesure permettrait de réaliser plusieurs bases de données liées au niveau de dégradation de la maladie. Ces bases pouvant être augmentées et utilisées comme des épisodes (notion vue dans notre revue des méthodes de « few-shot » en chapitre 4) pour créer un modèle par niveau de dégradation (et/ou type de maladie). De plus, l'utilisation de modèles auto-supervisés dans l'architecture d'une approche « few-shot » peut grandement faciliter la tâche. Pour se faire, nous voyons deux utilisations possibles :

— l'utilisation d'un modèle auto-supervisé (avec un peaufinage de ses paramètres sur l'ensemble d'entraînement de chaque épisode) comme entrée de l'architecture « few-shot ». Ainsi, le modèle auto-supervisé serait utilisé comme encodeur de parole.
— l'utilisation d'un modèle auto-supervisé comme base de l'architecture d'une approche « few-shot » (avec un peaufinage de ses paramètres sur l'ensemble d'entraînement de chaque épisode). Il sera alors nécessaire d'ajouter des couches permettant de réaliser les prédictions (en commençant, par exemple, à ajouter deux couches non linéaires). Ainsi, le modèle auto-supervisé serait utilisé pour réaliser un apprentissage par transfert avec une approche « few-shot ».

Dans les deux cas, le choix de la méthode « few-shot » choisie dépendra du nombre de classes à prédire (nombre de mots, d'émotions, de locuteurs) [41].

Nous pouvons également imaginer que les mesures de dégradations permettent la détection de maladie à travers certains symptômes de production de la parole. Cette détection pourrait être appliquée dans les assistants vocaux pour couvrir un maximum de personnes (il faudrait dans ce cas bien réfléchir à la protection des données des utilisateurs pour éviter la divulgation de secrets médicaux). Pour arriver à cette fin, il serait nécessaire d'étudier les symptômes de chaque maladie ciblée. Pour éviter un maximum de biais, la création ou l'utilisation de bases suivant des critères définis par [Pommée et al., 2022] nous semble être un minimum nécessaire. Ensuite, une vérité terrain d'experts sur le niveau de dégradation de chaque participant (patients et contrôles) est un prérequis minimum. Il serait par ailleurs important d'avoir une distribution homogène des patients. Cela nous paraît plus réalisable, car ce genre d'outils serait destiné à des personnes n'ayant pas déjà été diagnostiquées et donc n'étant pas gravement atteintes par la maladie (nous avons vu dans notre corpus que ce sont les participants les moins bien représentés). Ici encore, nous pourrions envisager l'utilisation d'approches « few-shot » associées à des modèles auto-supervisés pré-appris sur de grands corpus comme vue dans le paragraphe précédent. La différence ici serait que la prédiction se ferait soit sur une classe (parole saine/pathologique) soit sur le type de maladie et le niveau

41. Nous donnons des indications sur ces quantités dans le chapitre 4.

de dégradation. Dans le dernier cas, il est nécessaire d'avoir la vérité terrain associée et l'utilisation de fonction de coût multi-labels peut être envisagée (au lieu d'apprendre deux modèles distincts pour chaque sortie).

Par la similitude des contraintes dans le domaine du traitement automatique des langues sous-ressourcées, il serait également intéressant d'envisager une transposition des méthodes que j'ai proposées dans cette thèse. En particulier les méthodes de « few-shot » associées à des modèles auto-supervisés, dans le cadre de la reconnaissance de la parole pour les langues peu dotées en ressources.

A

Documentation de l'application livrée au CHU de Toulouse

A.1 Présentation de l'application livrée au CHU de Toulouse

Voici la présentation qui a été produite pour présenter l'application du Protopitch SAMI installée au CHU de Toulouse.

Application ProtoPitch

Sévèr'io

Sommaire

Contexte

Contexte applicatif

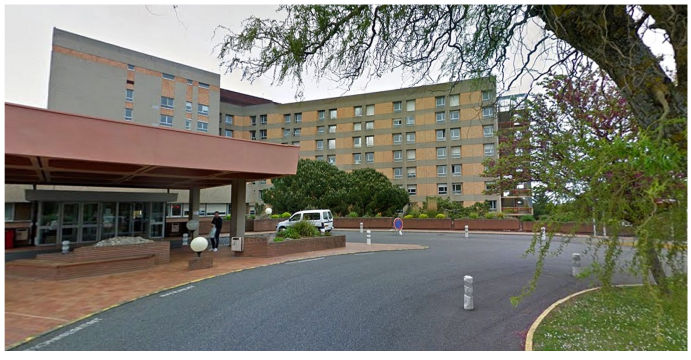


Illustration 1 - Hôpital Larrey

- Application pour le suivi ORL
- Mesure de la sévérité du trouble de la parole de patients atteints de cancer ORL
- Exemples d'utilisations
  - Voir l'impact de la rééducation sur le patient
  - Mesure l'évolution d'un patient

Fonctionnalités

Mesure d'intelligibilité

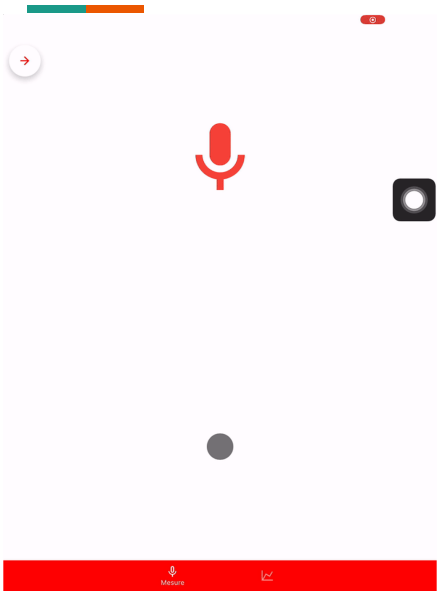


Illustration 2 - Séquence de mesure

- Permet une mesure d'intelligibilité en utilisant l'algorithme de Vincent (mesure entropique et Pase+)
- **Reste à intégrer** : Mise à l'échelle du score

Gestion des patients

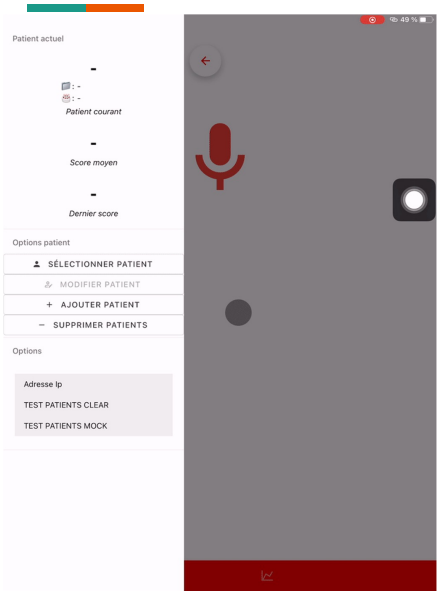


Illustration 3 - Gestion des patients

- Permet d'ajouter, supprimer, modifier les patients
- Permet de sélectionner le "**patient courant**"

Historique

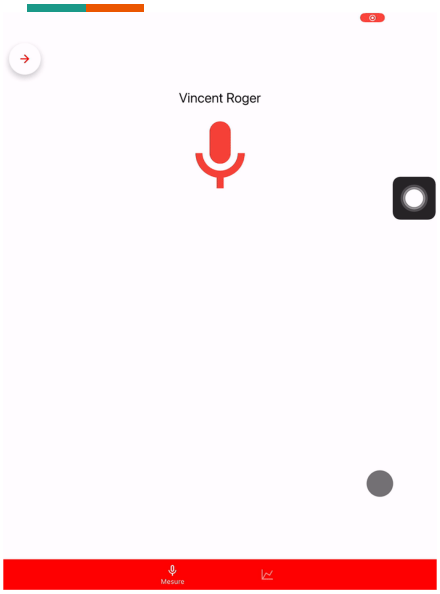


Illustration 4 - Historique

- Affichage de l'historique du **patient courant**
- Possibilité de trier par date / score
- Possibilité de mettre des filtres sur l'historique
- Historique crypté en AES 256

Technologies utilisées

Technologies de l'application

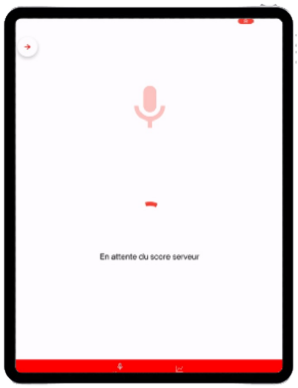


Illustration 5 - Technologies utilisées dans l'application

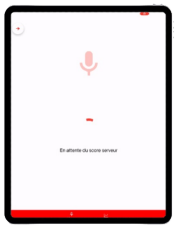
Technologies de l'application - React Native

React Native

- Permet de faire des applications IOS & Android
- Exemple d'applications : Facebook, Instagram, Airbnb, Skype...

Illustration 5 - Technologies utilisées dans l'application

Interactions Serveur



Application

Serveur

Illustration 8 - Interactions serveur & Technologies utilisées

Interactions Serveur

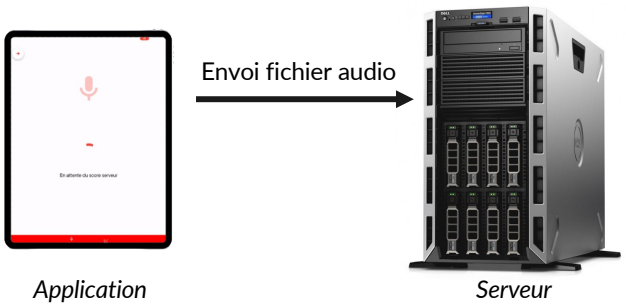


Illustration 8 - Interactions serveur & Technologies utilisées

Interactions Serveur

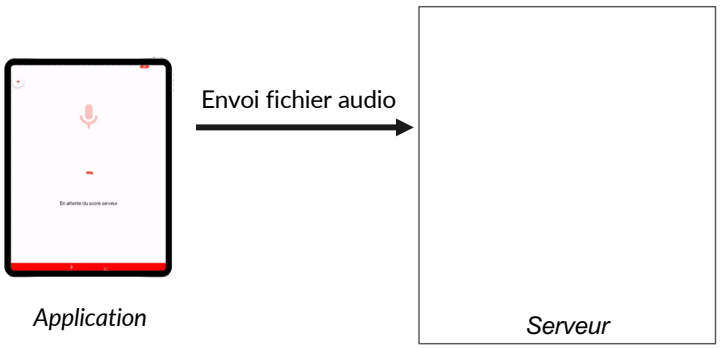
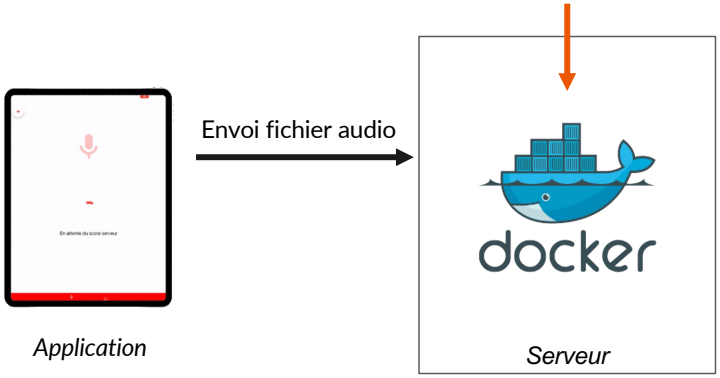


Illustration 8 - Interactions serveur & Technologies utilisées

Interactions Serveur - Docker



- Logiciel libre permettant de lancer des applications dans des “conteneurs logiciels”
  - Permet “d’empaqueter” une application (et ses dépendances) dans un conteneur
  - Permet de déployer facilement les algorithmes sur un ordinateur cible

Illustration 8 - Interactions serveur & Technologies utilisées

Interactions Serveur

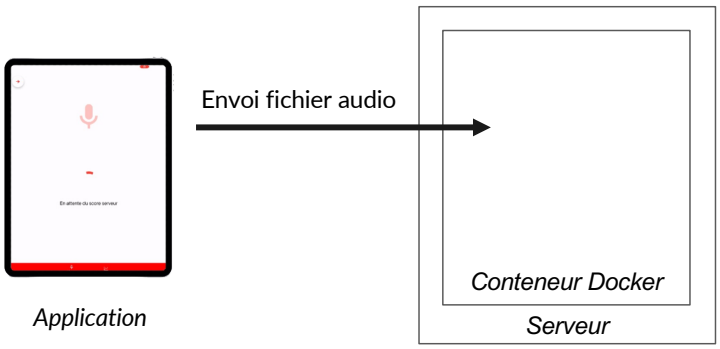
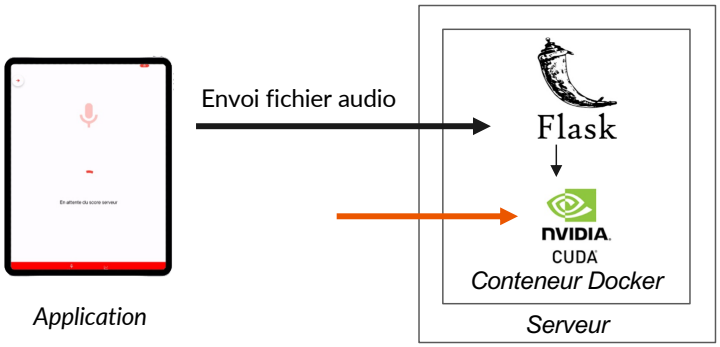


Illustration 8 - Interactions serveur & Technologies utilisées

Interactions Serveur - Cuda



- Le serveur gère les requêtes de l'application
- Il lance les algorithmes nécessitant du GPU

Illustration 8 - Interactions serveur & Technologies utilisées

Interactions Serveur

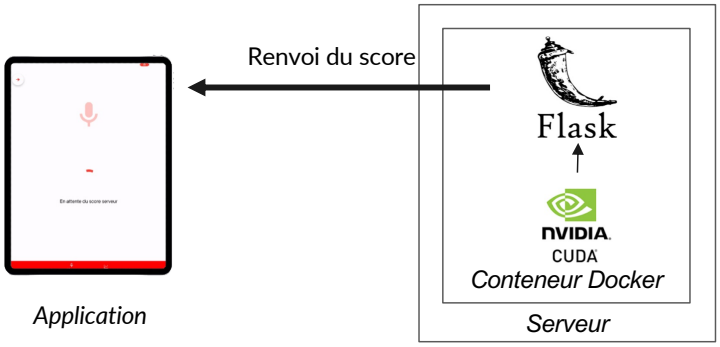


Illustration 8 - Interactions serveur & Technologies utilisées

“Bonus”

Documentation

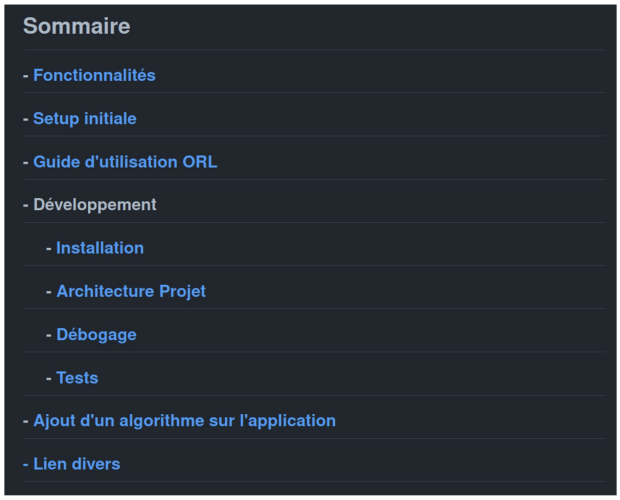


Illustration 9 - Documentation

- Documentation complète (sous forme de Markdown) disponible sur le git du projet

Tests

Illustration 10 - Tests

- Utilisation des librairies **Jest** & **Enzyme** pour faire des tests
- Utilisation d'un environnement iOS et Android
- **Note** : Problème sur une librairie d'input **JavaScript**

Mode "Plusieurs modes de mesures"

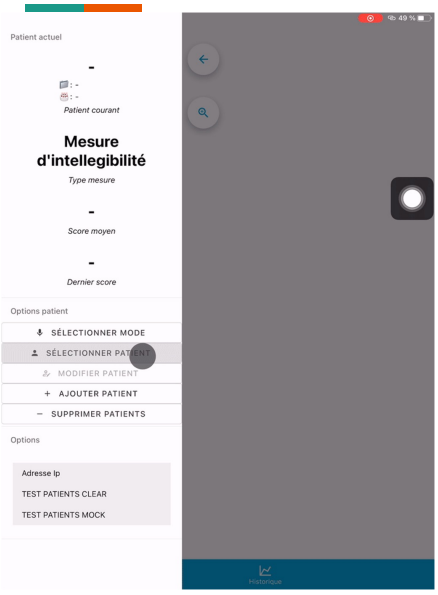


Illustration 11 - Mode plusieurs types de mesures

- Permet l'ajout de plusieurs modes de mesures
- Historique unique à chaque mode
- Ajout simplifié (uniquement une ligne à rajouter sur l'application)

A.2 Documentation de l'application

La documentation réalisée pour l'application du Protopitch SAMI se trouve sur le serveur gitlab suivant : https://gitlab.irit.fr/samova/documentation_protopitch. Elle comporte une partie destinée aux praticiens, des détails techniques sur l'application, dont les possibilités d'augmentations par de nouvelles mesures.

B

Détail de la réglementation RGPD appliquée

Ce travail de doctorat a été réalisé en respectant le cadre réglementaire du projet ANR RUGBI. Vous trouverez ci-dessous le détail des traitements réalisés dans ce cadre. Le numéro d'agrément de la CPP « Ouest IV » est le RC 31-18-0458 a été obtenu le 19 février 2020 et concerne en particulier les nouvelles données qui ont été acquises dans le cadre de ce projet. Le protocole ci-dessous a fait l'objet d'une déclaration au DPO CNRS par le biais de l'IRIT. Ces données font partie du Groupement d'Intérêt Scientifique PARALOTHEQUE [42], qui en assure la pérennité et l'accès aux partenaires scientifiques.

B.1 Cadre de l'étude

Produire des analyses et des modélisations automatiques basées sur les fichiers audio et les métadonnées visant à prédire l'intelligibilité ou localiser les zones les plus porteuses d'intelligibilité.

B.2 Données utilisées

- sous corpus provenant du projet INCA C2SI : enregistrements audio et métadonnées pseudo-anonymisées de patients atteints de cancer de la tête et du cou
- sous corpus provenant du projet PARK360 : enregistrements audio et métadonnées pseudo-anonymisées de patients atteints de la maladie de Parkinson
- corpus en cours de réalisation au CHU Toulouse du même type que le projet INCA C2SI mais présentant un suivi sur plusieurs années des mêmes patients (suivi longitudinal).

Les données sont enregistrées au CHU Toulouse, puis pseudo-anonymisées avant d'être versées sur le serveur OSIRIM (projet RUGBI).

Détail des données présentes :

- Enregistrements audio, format wave (48 kHz, 16 bits) (objet principal des études)

42. https://www.irit.fr/SAMOVA/site/projects/others/parolotheque/

— Métadonnées : index temporels des événements enregistrés (permet de localiser les différents mots prononcés, réponses apportées)
— Code identifiant unique du locuteur : code généré par le CHU pour identifier chaque personne
— Âge lors de l'enregistrement (la voix évoluant avec l'âge, il est utile de connaître l'âge pour prendre en compte les modifications apportées par la vieillesse)
— Genre (la voix est très dépendante du genre, les traitements doivent en tenir compte)
— Département de naissance (pour avoir une information sur un potentiel accent dans la voix)
— Département de résidence (pour avoir une information sur un potentiel accent dans la voix)
— Indicateur TNM : classification de la grosseur de la tumeur (l'altération de la voix est fortement dépendante de la taille de la tumeur)
— Antécédents chirurgie ou radiothérapie : très grande incidence sur la qualité de la voix
— Localisation tumeur : très grande incidence sur la qualité de la voix à traiter
— Résultats chiffrés de questionnaires sur la qualité de la vie (SHI, PHI, SF36) : indicateurs permettent d'analyser les corrélations entre la qualité de la voix et la qualité de la vie de la personne
— Résultats des évaluations perceptives (sévérité, intelligibilité) réalisées par un groupe d'experts au CHU : objectifs primaires des modélisations automatiques à réaliser.

B.3 Accès aux données

Le corpus cancer pseudo-anonymisé stocké sur OSIRIM (salle serveur non accessible au public, sécurisation des accès, log des connexions et des transferts de données).

Travail sur les données :
— Soit directement sur le serveur OSIRIM
— Soit copie des fichiers sur les ordinateurs cryptés des chercheurs (afin de réaliser des annotations supplémentaires, écouter les fichiers audio)

Liste des laboratoires ayant accès : LPL, LIA, UT2J-LNPL, IRIT.

Les travaux de ces laboratoires générant des métadonnées sont versés sur le serveur OSIRIM.

Bibliographie

[Abásolo et al., 2015] Abásolo, D., Simons, S., Morgado da Silva, R., Tononi, G., and Vyazovskiy, V. V. (2015). Lempel-ziv complexity of cortical activity during sleep and waking in rats. *Journal of neurophysiology*, 113(7) :2742–2752.

[Abderrazek et al., 2020] Abderrazek, S., Fredouille, C., Ghio, A., Lalain, M., Meunier, C., and Woisard, V. (2020). Towards Interpreting Deep Learning Models to Understand Loss of Speech Intelligibility in Speech Disorders — Step 1 : CNN Model-Based Phone Classification. In *Proc. Interspeech 2020*, pages 2522–2526.

[Anand et al., 2019] Anand, P., Singh, A. K., Srivastava, S., and Lall, B. (2019). Few shot speaker recognition using deep neural networks.

[Ardila et al., 2020] Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common voice : A massively-multilingual speech corpus. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., and Piperidis, S., editors, *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pages 4218–4222. European Language Resources Association.

[Auzou et al., 2007] Auzou, P., Auzou, P., Rolland Monnoury, V., Pinto, S., and Öszancak, C. (2007). Les objectifs du bilan de la dysarthrie. *Les dysarthries. Auzou P, Rolland Monnoury V, Pinto S, Öszancak C. Marseille : Solal*, pages 189–195.

[Baevski et al., 2020] Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0 : A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33 :12449–12460.

[Balaguer et al., 2019] Balaguer, M., Farinas, J., Pinquier, J., and Woisard, V. (2019). Construction of the automatic Carcinologic Speech Severity Index (C2SI) score. In *31st World Congress of the International Association of Logopedics and Phoniatrics (IALP)*, pages 1–15. IALP : International Association of Logopedics and Phoniatrics.

[Balaguer et al., 2020] Balaguer, M., Pommée, T., Farinas, J., Pinquier, J., Woisard, V., and Speyer, R. (2020). Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis : Systematic review. *Head & neck*, 42(1) :111–130.

[Barker et al., 2018] Barker, J., Watanabe, S., Vincent, E., and Trmal, J. (2018). The Fifth 'CHiME' Speech Separation and Recognition Challenge : Dataset, Task and Baselines. In *Interspeech 2018*, pages 1561–1565. ISCA.

[Benesty et al., 2009] Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

[Besacier et al., 2014] Besacier, L., Barnard, E., Karpov, A., and Schultz, T. (2014). Automatic speech recognition for under-resourced languages : A survey. *Speech Communication*, 56 :85–100.

[Beukelman and Yorkston, 1980] Beukelman, D. R. and Yorkston, K. M. (1980). Influence of passage familiarity on intelligibility estimates of dysarthric speech. *Journal of Communication Disorders*, 13(1) :33–41.

[Borggreven et al., 2007] Borggreven, P. A., Aaronson, N. K., Verdonck-de Leeuw, I. M., Muller, M. J., Heiligers, M. L., de Bree, R., Langendijk, J. A., and Leemans, C. R. (2007). Quality of life after surgical treatment for oral and oropharyngeal cancer : a prospective longitudinal assessment of patients reconstructed by a microvascular flap. *Oral oncology*, 43(10) :1034–1042.

[Brierley, 2016] Brierley, J. D. (2016). Mkge, christian wittekind (editor). tnm classification of malignant tumours.

[Busso et al., 2008] Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). IEMOCAP : interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42(4) :335–359.

[Chatziagapi et al., 2019] Chatziagapi, A., Paraskevopoulos, G., Sgouropoulos, D., Pantazopoulos, G., Nikandrou, M., Giannakopoulos, T., Katsamanis, A., Potamianos, A., and Narayanan, S. (2019). Data Augmentation Using GANs for Speech Emotion Recognition. In *Interspeech 2019*, pages 171–175. ISCA.

[Chen et al., 2019] Chen, L.-W., Lee, H.-Y., and Tsao, Y. (2019). Generative Adversarial Networks for Unpaired Voice Transformation on Impaired Speech. In *Interspeech 2019*, pages 719–723. ISCA.

[Chung et al., 2018] Chung, J. S., Nagrani, A., and Zisserman, A. (2018). VoxCeleb2 : Deep Speaker Recognition. In *Interspeech 2018*, pages 1086–1090. ISCA.

[Chung and Glass, 2018] Chung, Y.-A. and Glass, J. (2018). Speech2Vec : A Sequence-to-Sequence Framework for Learning Word Embeddings from Speech. In *Interspeech 2018*, pages 811–815. ISCA.

[Chung et al., 2021] Chung, Y.-A., Zhang, Y., Han, W., Chiu, C.-C., Qin, J., Pang, R., and Wu, Y. (2021). w2v-bert : Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250.

[Clapham et al., 2012] Clapham, R. P., van der Molen, L., van Son, R., van den Brekel, M. W., Hilgers, F. J., et al. (2012). Nki-ccrt corpus-speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy. In *LREC*, volume 4, pages 3350–3355. Citeseer.

[Deka et al., 2018] Deka, B., Chakraborty, J., Dey, A., Nath, S., Sarmah, P., Nirmala, S. R., and Vijaya, S. (2018). Speech corpora of under resourced languages of north-east india. In *2018 Oriental COCOSDA - International Conference on Speech Database and Assessments, Miyazaki, Japan, May 7-8, 2018*, pages 72–77.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT : Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of*

the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[Ding et al., 2020] Ding, S., Chen, T., Gong, X., Zha, W., and Wang, Z. (2020). Autospeech : Neural architecture search for speaker recognition. In Meng, H., Xu, B., and Zheng, T. F., editors, *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 916–920. ISCA.

[Dubnov, 2004] Dubnov, S. (2004). Generalization of spectral flatness measure for non-gaussian linear processes. *IEEE Signal Processing Letters*, 11(8) :698–701.

[Dwivedi et al., 2011] Dwivedi, R. C., St. Rose, S., Roe, J. W., Chisholm, E., Elmiyeh, B., Nutting, C. M., Clarke, P. M., Kerawala, C. J., Rhys-Evans, P. H., Harrington, K. J., et al. (2011). First report on the reliability and validity of speech handicap index in native english-speaking patients with head and neck cancer. *Head & neck*, 33(3) :341–348.

[Ellington, 2020] Ellington, T. D. (2020). Trends in incidence of cancers of the oral cavity and pharynx—united states 2007–2016. *MMWR. Morbidity and Mortality Weekly Report*, 69.

[Eloff et al., 2019] Eloff, R., Engelbrecht, H. A., and Kamper, H. (2019). Multimodal One-shot Learning of Speech and Images. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8623–8627.

[Enderby, 2013] Enderby, P. (2013). Disorders of communication : Dysarthria. In *Handbook of Clinical Neurology*, volume 110, pages 273–281. Elsevier.

[Falk et al., 2012] Falk, T. H., Chan, W.-Y., and Shein, F. (2012). Characterization of atypical vocal source excitation, temporal dynamics and prosody for objective measurement of dysarthric word intelligibility. *Speech Communication*, 54(5) :622–631.

[Fang et al., 2017] Fang, C., Li, H., Ma, L., and Zhang, M. (2017). Intelligibility Evaluation of Pathological Speech through Multigranularity Feature Extraction and Optimization. *Computational and Mathematical Methods in Medicine*, 2017 :1–8.

[Feng and Chaspari, 2021] Feng, K. and Chaspari, T. (2021). Few-shot learning in emotion recognition of spontaneous speech using a siamese neural network with adaptive sample pair formation. *IEEE Transactions on Affective Computing*.

[Fletcher et al., 2017] Fletcher, A. R., Wisler, A. A., McAuliffe, M. J., Lansford, K. L., and Liss, J. M. (2017). Predicting intelligibility gains in dysarthria through automated speech feature analysis. *Journal of Speech, Language, and Hearing Research*, 60(11) :3058–3068.

[Fougeron et al., 2010] Fougeron, C., Crevier-Buchman, L., Fredouille, C., Ghio, A., Meunier, C., Chevrie-Muller, C., Audibert, N., Bonastre, J.-F., Colazo-Simon, A., Delooze, C., et al. (2010). Developing an acoustic-phonetic characterization of dysarthric speech in french. In *7th International Conference on Language Resources, Technologies and Evaluation (LREC)*, pages 2831–2838. Nicoletta Calzolari (Conference Chair) and Khalid Choukri and Bente Maegaard ….

[Garcia and Bruna, 2018] Garcia, V. and Bruna, J. (2018). Few-Shot Learning with Graph Neural Networks. In *ICLR 2018*, page 13.

[Garofolo et al., 1993] Garofolo, J. S., Lamel, L. F., Fisher, W. M., Fiscus, J. G., and Pallett, D. S. (1993). Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93 :27403.

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

[Ghio et al., 2012] Ghio, A., Pouchoulin, G., Teston, B., Pinto, S., Fredouille, C., De Looze, C., Robert, D., Viallet, F., and Giovanni, A. (2012). How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers ? *Speech Communication*, 54(5) :664–679. Advanced Voice Function Assessment.

[Godfrey et al., 1992] Godfrey, J. J., Holliman, E. C., and McDaniel, J. (1992). Switchboard : telephone speech corpus for research and development. In *[Proceedings] ICASSP-92 : 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520.

[Hernandez et al., 2018] Hernandez, F., Nguyen, V., Ghannay, S., Tomashenko, N., and Estève, Y. (2018). TED-LIUM 3 : Twice as much data and corpus repartition for experiments on speaker adaptation. In Karpov, A., Jokisch, O., and Potapova, R., editors, *Speech and Computer - 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18-22, 2018, Proceedings*, volume 11096 of *Lecture Notes in Computer Science*, pages 198–208. Springer.

[Hozjan et al., 2002] Hozjan, V., Kacic, Z., Moreno, A., Bonafonte, A., and Nogueiras, A. (2002). Interface databases : Design and collection of a multilingual emotional speech database. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02)*, Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA).

[Hsu et al., 2021] Hsu, W.-N., Bolte, B., Tsai, Y.-H. H., Lakhotia, K., Salakhutdinov, R., and Mohamed, A. (2021). Hubert : Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29 :3451–3460.

[Ioffe and Szegedy, 2015] Ioffe, S. and Szegedy, C. (2015). Batch normalization : Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR.

[Jacobi et al., 2010] Jacobi, I., van der Molen, L., Huiskens, H., Van Rossum, M. A., and Hilgers, F. J. (2010). Voice and speech outcomes of chemoradiation for advanced head and neck cancer : a systematic review. *European Archives of Oto-Rhino-Laryngology*, 267(10) :1495–1505.

[Janbakhshi et al., 2019] Janbakhshi, P., Kodrasi, I., and Bourlard, H. (2019). Spectral subspace analysis for automatic assessment of pathological speech intelligibility. In *Proc. Interspeech 2019*, pages 3038–3042.

[Jensen and Taal, 2016] Jensen, J. and Taal, C. H. (2016). An algorithm for predicting the intelligibility of speech masked by modulated noise maskers. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11) :2009–2022.

[Jiao et al., 2018] Jiao, Y., Tu, M., Berisha, V., and Liss, J. (2018). Simulating dysarthric speech for training data augmentation in clinical speech applications. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6009–6013.

[Kahn et al., 2020] Kahn, J., Riviere, M., Zheng, W., Kharitonov, E., Xu, Q., Mazaré, P.-E., Karadayi, J., Liptchinsky, V., Collobert, R., Fuegen, C., et al. (2020). Libri-light : A benchmark for asr with limited or no supervision. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. IEEE.

[Kent, 1992] Kent, R. D. (1992). *Intelligibility in speech disorders : Theory, measurement and management*, volume 1. John Benjamins Publishing.

[Kim et al., 2019] Kim, C., Shin, M., Garg, A., and Gowda, D. (2019). Improved Vocal Tract Length Perturbation for a State-of-the-Art End-to-End Speech Recognition System. In *Interspeech 2019*, pages 739–743. ISCA.

[Kim et al., 2008] Kim, H., Hasegawa-Johnson, M., Perlman, A., Gunderson, J., Huang, T. S., Watkin, K., and Frame, S. (2008). Dysarthric speech database for universal access research. In *Ninth Annual Conference of the International Speech Communication Association*.

[Koch et al., 2015] Koch, G., Zemel, R., and Salakhutdinov, R. (2015). Siamese Neural Networks for One-shot Image Recognition. *ICML Deep Learning Workshop*, page 8.

[Laaridh et al., 2017] Laaridh, I., Kheder, W. B., Fredouille, C., and Meunier, C. (2017). Automatic Prediction of Speech Evaluation Metrics for Dysarthric Speech. In *Interspeech*, pages 1834–1838, Stockholm, Sweden.

[Lamel et al., 1991] Lamel, L. F., Gauvain, J.-L., Eskénazi, M., et al. (1991). Bref, a large vocabulary spoken corpus for french1. *training*, 22(28) :50.

[Latif et al., 2020] Latif, S., Qadir, J., Qayyum, A., Usama, M., and Younis, S. (2020). Speech technology for healthcare : Opportunities, challenges, and state of the art. *IEEE Reviews in Biomedical Engineering*, 14 :342–356.

[Li and Abu-Mostafa, 2006] Li, L. and Abu-Mostafa, Y. S. (2006). Data complexity in machine learning. Computer Science Technical Report CaltechCSTR :2006.004, California Institute of Technology, Pasadena, USA.

[Li et al., 2019] Li, Y., Zhao, T., and Kawahara, T. (2019). Improved End-to-End Speech Emotion Recognition Using Self Attention Mechanism and Multitask Learning. In *Interspeech 2019*, pages 2803–2807. ISCA.

[Lian et al., 2020] Lian, Z., Tao, J., Liu, B., Huang, J., Yang, Z., and Li, R. (2020). Context-Dependent Domain Adversarial Neural Network for Multimodal Emotion Recognition. In *Proc. Interspeech 2020*, pages 394–398.

[Lüscher et al., 2019] Lüscher, C., Beck, E., Irie, K., Kitza, M., Michel, W., Zeyer, A., Schlüter, R., and Ney, H. (2019). RWTH ASR Systems for LibriSpeech : Hybrid vs Attention. In *Interspeech 2019*, pages 231–235. ISCA.

[Maier et al., 2009] Maier, A., Haderlein, T., Stelzle, F., Nöth, E., Nkenke, E., Rosanowski, F., Schützenberger, A., and Schuster, M. (2009). Automatic speech recognition systems for the evaluation of voice and speech disorders in head and neck cancer. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010 :1–7.

[Martínez et al., 2012] Martínez, D., Lleida, E., Ortega, A., Miguel, A., and Villalba, J. (2012). Voice pathology detection on the saarbrücken voice database with calibration and fusion of scores using multifocal toolkit. In *Advances in Speech and Language Technologies for Iberian Languages*, pages 99–109. Springer.

[Martins et al., 2015] Martins, P. C., Couto, T. E., and Gama, A. C. C. (2015). Auditory-perceptual evaluation of the degree of vocal deviation : correlation between the visual analogue scale and numerical scale. In *Codas*, volume 27, pages 279–284. SciELO Brasil.

[McFee et al., 2015] McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa : Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, pages 18–25. Citeseer.

[Middag et al., 2008] Middag, C., Van Nuffelen, G., Martens, J.-P., and De Bodt, M. (2008). Objective intelligibility assessment of pathological speakers. In *9th annual conference of the international speech communication association (interspeech 2008)*, pages 1745–1748. International Speech Communication Association (ISCA).

[Mirza and Osindero, 2014] Mirza, M. and Osindero, S. (2014). Conditional generative adversarial nets.

[Moattar and Homayounpour, 2009] Moattar, M. H. and Homayounpour, M. M. (2009). A simple but efficient real-time Voice Activity Detection algorithm. In *17th European Signal Processing Conference*, pages 2549–2553.

[Moore et al., 2018] Moore, M., Venkateswara, H., and Panchanathan, S. (2018). Whistle-blowing ASRs : Evaluating the Need for More Inclusive Speech Recognition Systems. In *Interspeech 2018*, pages 466–470. ISCA.

[Mustafa et al., 2014] Mustafa, M. B., Salim, S. S., Mohamed, N., Al-Qatab, B., and Siong, C. E. (2014). Severity-Based Adaptation with Limited Data for ASR to Aid Dysarthric Speakers. *PLoS ONE*, 9(1) :e86285.

[Nagrani et al., 2020] Nagrani, A., Chung, J. S., Xie, W., and Zisserman, A. (2020). Voxceleb : Large-scale speaker verification in the wild. *Computer Speech & Language*, 60 :101027.

[Nagrani et al., 2017] Nagrani, A., Chung, J. S., and Zisserman, A. (2017). VoxCeleb : A Large-Scale Speaker Identification Dataset. In *Interspeech 2017*, pages 2616–2620. ISCA.

[Nguyen et al., 2020] Nguyen, T.-S., Stueker, S., Niehues, J., and Waibel, A. (2020). Improving sequence-to-sequence speech recognition training with on-the-fly data augmentation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7689–7693. IEEE.

[Oord et al., 2018] Oord, A. v. d., Li, Y., and Vinyals, O. (2018). Representation Learning with Contrastive Predictive Coding. *CoRR*.

[Panayotov et al., 2015] Panayotov, V., Chen, G., Povey, D., and Khudanpur, S. (2015). Librispeech : An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, South Brisbane, Queensland, Australia.

[Parcollet et al., 2018] Parcollet, T., Ravanelli, M., Morchid, M., Linarès, G., and De Mori, R. (2018). Speech recognition with quaternion neural networks. In *NeurIPS 2018 - IRASL*.

[Park et al., 2019] Park, D. S., Chan, W., Zhang, Y., Chiu, C.-C., Zoph, B., Cubuk, E. D., and Le, Q. V. (2019). Specaugment : A simple data augmentation method for automatic speech recognition. In Kubin, G. and Kacic, Z., editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 2613–2617. ISCA.

[Pascual et al., 2019] Pascual, S., Ravanelli, M., Serrà, J., Bonafonte, A., and Bengio, Y. (2019). Learning Problem-Agnostic Speech Representations from Multiple Self-Supervised Tasks. In *Interspeech 2019*, pages 161–165. ISCA.

[Pernon et al., 2020] Pernon, M., Levêque, N., Delvaux, V., Assal, F., Borel, S., Fougeron, C., Trouville, R., and Laganaro, M. (2020). Monpage, un outil de screening francophone informatise d'évaluation perceptive et acoustique des troubles moteurs de la parole (dysarthries, apraxie de la parole). *Rééducation orthophonique*, 281 :171–97.

[Pommée et al., 2021] Pommée, T., Balaguer, M., Mauclair, J., Pinquier, J., and Woisard, V. (2021). Intelligibility and comprehensibility : A delphi consensus study. *International Journal of Language & Communication Disorders*.

[Pommée et al., 2022] Pommée, T., Balaguer, M., Mauclair, J., Pinquier, J., and Woisard, V. (2022). Criteria for creating new standard reading passages for the assessment of speech and voice : A delphi consensus study. *Clinical Linguistics & Phonetics*, pages 1–20.

[Quintas et al., 2020] Quintas, S., Mauclair, J., Woisard, V., and Pinquier, J. (2020). Automatic Prediction of Speech Intelligibility Based on X-Vectors in the Context of Head and Neck Cancer. In *Proc. Interspeech 2020*, pages 4976–4980.

[Ravanelli and Bengio, 2018] Ravanelli, M. and Bengio, Y. (2018). Interpretable Convolutional Filters with SincNet. In *NIPS 2018 Workshop IRASL*.

[Ravanelli et al., 2018] Ravanelli, M., Brakel, P., Omologo, M., and Bengio, Y. (2018). Light gated recurrent units for speech recognition. *IEEE Trans. Emerg. Top. Comput. Intell.*, 2(2) :92–102.

[Ravanelli et al., 2019] Ravanelli, M., Parcollet, T., and Bengio, Y. (2019). The pytorch-kaldi speech recognition toolkit. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6465–6469.

[Ravanelli et al., 2020] Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J., and Bengio, Y. (2020). Multi-task self-supervised learning for Robust Speech Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6989–6993. IEEE.

[Ravi and Larochelle, 2017] Ravi, S. and Larochelle, H. (2017). Optimization as a Model for Few-Shot Learning. In *ICLR 2017*, page 11. OpenReview.net.

[Roger et al., 2022a] Roger, V., Farinas, J., and Pinquier, J. (2022a). Deep neural networks for automatic speech processing : a survey from large corpora to limited data. *EURASIP*. Received : 19 November 2021 / Accepted : 15 July 2022.

[Roger et al., 2022b] Roger, V., Farinas, J., Woisard, V., and Pinquier, J. (2022b). Création d'une mesure entropique de la parole pour évaluer l'intelligibilité de patients atteints de cancers des voies aérodigestives supérieures. In *34e Journées d'Études sur la Parole (JEP2022)*, page A paraître, Noirmoutier, France. Association Française de la Communication Parlée. A paraître.

[Sahu et al., 2018] Sahu, P., Dua, M., and Kumar, A. (2018). Challenges and issues in adopting speech recognition. In *Speech and language processing for human-machine communications*, pages 209–215. Springer.

[Salimans et al., 2016] Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved Techniques for Training GANs. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 2234–2242. Curran Associates, Inc.

[Schneider et al., 2019] Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec : Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, pages 3465–3469.

[Shor et al., 2019] Shor, J., Emanuel, D., Lang, O., Tuval, O., Brenner, M., Cattiau, J., Vieira, F., McNally, M., Charbonneau, T., Nollstadt, M., Hassidim, A., and Matias, Y. (2019). Personalizing ASR for Dysarthric and Accented Speech with Limited Data. In *Interspeech 2019*, pages 784–788. ISCA.

[Snell et al., 2017] Snell, J., Swersky, K., and Zemel, R. (2017). Prototypical Networks for Few-shot Learning. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, volume 30, pages 4077–4087. Curran Associates, Inc.

[Szegedy et al., 2016] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Thomas et al., 2009] Thomas, L., Jones, T. M., Tandon, S., Carding, P., Lowe, D., and Rogers, S. (2009). Speech and voice outcomes in oropharyngeal cancer and evaluation of the university of washington quality of life speech domain. *Clinical Otolaryngology*, 34(1) :34–42.

[Van der Maaten and Hinton, 2008] Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

[Vinyals et al., 2016] Vinyals, O., Blundell, C., Lillicrap, T., kavukcuoglu, k., and Wierstra, D. (2016). Matching Networks for One Shot Learning. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, volume 29, pages 3630–3638. Curran Associates, Inc.

[Walshe and Miller, 2011] Walshe, M. and Miller, N. (2011). Living with acquired dysarthria : The speaker's perspective. *Disability and Rehabilitation*, 33(3) :195–203.

[Wang et al., 2018] Wang, K., Zhang, J., Sun, S., Wang, Y., Xiang, F., and Xie, L. (2018). Investigating generative adversarial networks based speech dereverberation for robust speech recognition. In Yegnanarayana, B., editor, *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 1581–1585. ISCA.

[Woisard et al., 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., and Roger, V. (2021). C2si corpus : a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers. *Language Resources and Evaluation*, 55(1) :173–190.

[Woisard and Lepage, 2010] Woisard, V. and Lepage, B. (2010). Perception of speech disorders : Difference between the degree of intelligibility and the degree of severity. *Audiological Medicine*, 8 :171–178.

[Yamagishi et al., 2019] Yamagishi, J., Veaux, C., MacDonald, K., et al. (2019). Cstr vctk corpus : English multi-speaker corpus for cstr voice cloning toolkit (version 0.92). Technical report, University of Edinburgh. The Centre for Speech Technology Research (CSTR).

[Yamasaki et al., 2017] Yamasaki, R., Madazio, G., Leão, S. H., Padovani, M., Azevedo, R., and Behlau, M. (2017). Auditory-perceptual evaluation of normal and dysphonic voices using the voice deviation scale. *Journal of Voice*, 31(1) :67–71.

[Yang et al., 2021] Yang, S.-w., Chi, P.-H., Chuang, Y.-S., Lai, C.-I. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G.-T., Huang, T.-H., Tseng, W.-C., Lee, K.-t., Liu, D.-R., Huang, Z., Dong, S., Li, S.-W., Watanabe, S., Mohamed, A., and Lee, H.-y. (2021). SUPERB : Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

[Yosinski et al., 2014] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks ? In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. D., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

[Zhang et al., 2006] Zhang, S., Jiang, H., Zhang, S., and Xu, B. (2006). Fast SVM training based on the choice of effective samples for audio classification. In *Proc. Interspeech 2006*, pages paper 1073–Wed1FoP.1.

[Zhang et al., 2019] Zhang, S., Qin, Y., Sun, K., and Lin, Y. (2019). Few-Shot Audio Classification with Attentional Graph Neural Networks. In *Interspeech 2019*, pages 3649–3653. ISCA.

[Zhang et al., 2020] Zhang, Y., Qin, J., Park, D. S., Han, W., Chiu, C.-C., Pang, R., Le, Q. V., and Wu, Y. (2020). Pushing the limits of semi-supervised learning for automatic speech recognition. *CoRR*, abs/2010.10504.

[Zhao et al., 2019] Zhao, S., Ni, C., Tong, R., and Ma, B. (2019). Multi-Task Multi-Network Joint-Learning of Deep Residual Networks and Cycle-Consistency Generative Adversarial Networks for Robust Speech Recognition. In *Interspeech 2019*, pages 1238–1242. ISCA.

Résumé

Les personnes atteintes de cancers des voies aérodigestives supérieures présentent des difficultés de prononciation après des chirurgies ou des radiothérapies. Il est important pour le praticien de pouvoir disposer d'une mesure reflétant la sévérité de la parole. Pour produire cette mesure, il est communément pratiqué une étude perceptive qui rassemble un groupe de cinq à six experts cliniques. Ce procédé limite l'usage de cette évaluation en pratique. Ainsi, la création d'une mesure automatique, semblable à l'indice de sévérité, permettrait un meilleur suivi des patients en facilitant son obtention.

Pour réaliser une telle mesure, nous nous sommes appuyés sur une tâche de lecture, classiquement réalisée. Nous avons utilisé les enregistrements du corpus cancer qui rassemble plus de 100 personnes. Ce corpus représente environ une heure d'enregistrement pour modéliser l'indice de sévérité.

Dans ce travail de doctorat, une revue des méthodes de l'état de l'art sur la reconnaissance de la parole, des émotions et du locuteur utilisant peu de données a été entreprise. Nous avons ensuite essayé de modéliser la sévérité à l'aide d'apprentissage par transfert et par apprentissage profond. Les résultats étant non utilisables, nous nous sommes tourné sur les techniques dites « few shot » (apprentissage à partir de quelques exemples seulement). Ainsi, après de premiers essais prometteurs sur la reconnaissance de phonèmes, nous avons obtenu des résultats prometteurs pour catégoriser la sévérité des patients. Néanmoins, l'exploitation de ces résultats pour une application médicale demanderait des améliorations.

Nous avons donc réalisé des projections des données de notre corpus. Comme certaines tranches de scores étaient séparables à l'aide de paramètres acoustiques, nous avons proposé une nouvelle méthode de mesure entropique. Celle-ci est fondée sur des représentations de la parole autoapprise sur le corpus Librispeech : le modèle PASE+, qui est inspiré de l'Inception Score (généralement utilisé en image pour évaluer la qualité des images générées par les modèles). Notre méthode nous permet de produire un score semblable à l'indice de sévérité avec une corrélation de Spearman de 0,87 sur la tâche de lecture du corpus cancer. L'avantage de notre approche est qu'elle ne nécessite pas des données du corpus cancer pour l'apprentissage. Ainsi, nous pouvons utiliser l'entièreté du corpus pour l'évaluation de notre système. La qualité de nos résultats nous a permis d'envisager une utilisation en milieu clinique à travers une application sur tablette : des tests sont d'ailleurs en cours à l'hôpital Larrey de Toulouse.

Mots-clés: Parole pathologique, indice de sévérité, trouble de la parole, cancer ORL, apprentissage profond, apprentissage avec quelques exemples, auto supervisé, mesure entropique, few-shot, peu de données, quantité de données limités, traitement automatique de la parole.

Abstract

People with head and neck cancers have speech difficulties after surgery or radiation therapy. It is important for health practitioners to have a measure that reflects the severity of speech. To produce this measure, a perceptual study is commonly performed with a group of five to six clinical experts. This process limits the use of this assessment in practice. Thus, the creation of an automatic measure, similar to the severity index, would allow a better follow-up of the patients by facilitating its obtaining.

To realise such a measure, we relied on a reading task, classically performed. We used the recordings of the cancer corpus, which includes more than 100 people. This corpus represents about one hour of recording to model the severity index.

In this PhD work, a review of state-of-the-art methods on speech, emotion and speaker recognition using little data was undertaken. We then attempted to model severity using transfer learning and deep learning. Since the results were not usable, we turned to the so-called « few shot » techniques (learning from only a few examples). Thus, after promising first attempts at phoneme recognition, we obtained promising results for categorising the severity of patients. Nevertheless, the exploitation of these results for a medical application would require improvements.

We therefore performed projections of the data from our corpus. As some score slices were separable using acoustic parameters, we proposed a new entropic measurement method. This one is based on self-supervised speech representations on the Librispeech corpus : the PASE+ model, which is inspired by the Inception Score (generally used in image processing to evaluate the quality of images generated by models). Our method allows us to produce a score similar to the severity index with a Spearman correlation of 0.87 on the reading task of the cancer corpus. The advantage of our approach is that it does not require data from the cancer corpus for training. Thus, we can use the whole corpus for the evaluation of our system. The quality of our results has allowed us to consider a use in a clinical environment through an application on a tablet : tests are underway at the Larrey Hospital in Toulouse.

Keywords: Speech pathology, severity index, speech disorder, ENT cancer, deep learning, learning with a few examples, self-supervised, entropic measurement, few-shot, limited data, limited amount of data, automatic speech processing