Apprentissage profond pour l'analyse de la parole pathologique : étude comparative entre modèles CNN et à base de transformers

Malo Maisonneuve, Corinne Fredouille, Muriel Lalain, Alain Ghio, Virginie Woisard

Apprentissage profond pour l'analyse de la parole pathologique : étude comparative entre modèles CNN et à base de transformers

Malo Maisonneuve[1] Corinne Fredouille[1] Muriel Lalain[2] Alain Ghio[2]
Virginie Woisard[3]
(1) LIA, Avignon University, France
(2) Aix-Marseille University, CNRS, LPL, Aix-en-Provence, France
(3) LNPL, Toulouse University and Toulouse Hospital, Toulouse, France
prénom.nom@[univ-avignon ; univ-amu].fr

RÉSUMÉ

Les cancers des voies aérodigestives supérieures (VADS) ont un impact significatif sur la capacité des patients à s'exprimer, ce qui affecte leur qualité de vie. Les évaluations actuelles de la parole pathologique sont subjectives, justifiant le besoin de méthodes automatiques et objectives. Un modèle auto-supervisé basé sur Wav2Vec2 est proposé pour la classification de phonèmes chez les patients atteints de cancer des VADS, visant une amélioration des taux de bonne classification et une meilleure discrimination des caractéristiques phonétiques. Les impacts des paramètres d'affinage, des données de pré-entraînement, de la taille du modèle et des données d'affinage sont explorés. Nos résultats montrent que l'architecture Wav2Vec2 surpasse une approche basée sur un CNN, et montre une corrélation significative avec les mesures perceptives. Ce travail ouvre la voie à une meilleure compréhension de la parole pathologique, via une représentation auto-apprise de la parole, très pertinente pour des approches d'interprétation à destination des cliniciens.

ABSTRACT

Deep learning for speech pathology : a comparative analysis of CNN and transformer-based models

Head and neck cancers significantly impact patients' ability to speak, affecting their quality of life. Commonly used metrics for assessing pathological speech are subjective, prompting the need for automated and unbiased evaluation methods. This study proposes a self-supervised Wav2Vec2-based model for phone classification in HNC patients, aiming to enhance accuracy and improve the discrimination of phonetic features. The impact of fine-tuning parameters, pre-training datasets, model size, and fine-tuning datasets will be explored. Evaluation on diverse corpora reveals the effectiveness of the Wav2Vec2 architecture, outperforming a CNN-based approach, used in previous work. Correlation with perceptual measures also affirms the model's relevance for clinical speech analysis. This work paves the way for a better understanding of pathological speech, by leveraging a complex self-learnt speech representation, relevant for interpretability approaches for clinicians.

MOTS-CLÉS : troubles de la parole, cancer de la tête et du cou, apprentissage profond, classification de phonèmes, intelligibilité, interprétabilité.

KEYWORDS: speech disorders, Head and Neck Cancer, deep learning, phone classification, intelligibility, interpretability.

Traduction de l'article accepté à Interspeech 2024 : "Towards objective and interpretable speech disorder assessment: a comparative analysis of CNN and transformer-based models"

1 Introduction

Les cancers des voies aérodigestives supérieures (VADS) affectent les voies respiratoires et digestives supérieures, notamment la cavité buccale, le pharynx, le larynx, la cavité nasale et les glandes salivaires. Le traitement de ce cancer, que ce soit par radiothérapie, chimiothérapie et/ou chirurgie, peut impacter significativement la parole des patientes et patients qui en souffrent. La difficulté à communiquer avec les autres a ainsi un impact négatif sur leur qualité de vie. Évaluer correctement la parole qu'ils ou elles produisent, en identifiant le niveau d'altération et ce qui la rend atypique est essentiel pour les aider au mieux lors de séances de rééducation. Malheureusement, les métriques couramment utilisées telles que la sévérité ou le niveau d'intelligibilité sont subjectives et susceptibles d'être mal évaluées, même par des experts (Astésano *et al.*, 2018). Par ailleurs, ces métriques n'apportent aucune information, outre un score, sur la nature des dégradations mesurées. Proposer une manière automatique d'évaluer la parole pathologique est essentiel pour pouvoir baser les stratégies de rééducation sur des évaluations objectives et non biaisées.

Récemment, des modèles auto-supervisés ont montré leur succès dans la capture de concepts phonétiques et de diverses caractéristiques de la parole. Dans (tom Dieck *et al.*, 2022), les auteurs ont montré que les modèles Wav2Vec2 sont capables d'apprendre certains concepts phonétiques et qu'ils modélisent correctement le lieu et le mode d'articulation. Certaines recherches se sont concentrées sur l'utilisation de ces modèles pour évaluer automatiquement le niveau de sévérité de la parole (Hernandez *et al.*, 2022; Favaro *et al.*, 2023; Yeo *et al.*, 2023a; Javanmardi *et al.*, 2024). Bien que la détection d'une maladie et son évaluation quantitative soit importante, nous pensons qu'il est tout aussi important d'expliquer les résultats de ces modèles. En effet, leur explicabilité permet de cibler leurs faiblesses, et ainsi d'accroître la confiance que les cliniciens peuvent avoir en ces systèmes. Jusqu'à présent, un nombre limité d'études se sont concentrées sur l'interprétabilité de ces modèles. Dans (Tu *et al.*, 2017), un modèle prédictif de la sévérité de la parole dysarthrique a été entraîné, en incorporant une couche *bottleneck* dans un réseau neuronal à couches entièrement connectées pour améliorer l'interprétabilité. En effet, un apprentissage par transfert a été utilisé pour apprendre des étiquettes interprétables cliniquement issues de la classification de Darley (Darley *et al.*, 1969). Leurs résultats montrent une amélioration de la précision de l'évaluation de la dysarthrie, avec des justifications basées sur des caractéristiques interprétables. Une extension de ce travail (Xu *et al.*, 2023) a évité les étiquettes perceptives (nécessitant une annotation experte et coûteuse en temps des enregistrements utilisés pour l'entraînement et les tests des modèles) en formant la couche interprétable autour de quatre caractéristiques acoustiques liées à la dysarthrie et extraites automatiquement. L'utilisation de SHAP (*SHapley Additive exPlanations* (Lundberg & Lee, 2017)) a permis d'analyser la contribution de chaque caractéristique acoustique à la prédiction finale. Des avancées récentes ont aussi démontré l'efficacité de l'utilisation de la métrique *Goodness of Pronunciation*, en utilisant la partie d'extraction de caractéristiques – figée – du modèle Wav2Vec2, suivi d'un classifieur de phonèmes (Yeo *et al.*, 2023b). Leur approche montre l'impact relatif de chaque phonème sur le score prédit de sévérité. Une autre méthodologie a été utilisée dans (Abderrazek *et al.*, 2023) en utilisant un réseau neuronal convolutif (CNN) pour la classification de phonèmes. La méthode NCD (*Neuro-Concept Detector*) liée à la métrique ANPS (*Artificial Neuron-based Phonological Similarity*) permettent une interprétation des scores de sévérité et d'intelligibilité prédits, en terme d'altérations de la parole produite par des patients, via l'association des neurones du classifieur aux traits phonétiques qu'ils détectent.

Notre travail vise à étendre cette interprétation en remplaçant le CNN par un modèle auto-supervisé

basé sur l'architecture complète de Wav2Vec2. Cette modification a pour objectif de fournir de meilleurs taux de bonne classification, mais aussi d'apporter ultérieurement une description plus nuancée et plus riche des caractéristiques phonétiques de la parole pathologique. En plus du changement d'architecture du modèle, ce travail examinera l'impact du choix des corpus de pré-entraînement des modèles Wav2Vec2 disponibles à la communauté, de la taille de ces modèles, de l'étape d'affinage et des corpus utilisés lors de celle-ci sur le choix d'un modèle auto-supervisé pour la tâche visée de classification de phonèmes. Cette exploration est cruciale pour optimiser les performances du modèle sur des corpus variés, contribuant à la robustesse et à la capacité de généralisation de notre approche. Ainsi, une analyse détaillée des confusions entre phonèmes sera menée pour valider la capacité de généralisation de nos modèles sur d'autres ensembles de données. Ensuite, nous analyserons la corrélation entre les taux de bonne classification de phonèmes et les mesures perceptives obtenues auprès d'experts sur la parole de patients ayant bénéficié de soins suite à un cancer des VADS. Cette approche d'évaluation multidimensionnelle fournira une évaluation complète du modèle auto-supervisé proposé, démontrant son potentiel pour une analyse précise et cliniquement pertinente de la parole pathologique.

2 Corpus

Pour entraîner et évaluer nos modèles pendant l'entraînement, nous nous sommes appuyés sur un sous-ensemble de BREF (Lamel *et al.*, 1991) et de Common Phone (Klumpp *et al.*, 2022). Pour tester nos modèles une fois affinés, nous nous sommes basés sur BREF-Int et C2SI (Astésano *et al.*, 2018).
Le corpus **BREF** est un corpus de 120 locuteurs français lisant des extraits du journal *Le Monde*. Les enregistrements ont eu lieu dans les années 90, avec des personnes recrutées dans la région parisienne. Un sous-ensemble équilibré en termes de phonèmes de ce corpus a été créé pour garantir que les modèles affinés ne présentent pas de biais envers un phonème spécifique. Il est composé de plus de trois millions de trames de 127 ms, chacune centrée sur un phonème français ou un silence. Nous utilisons également l'ensemble de données **BREF-Int**, un sous-ensemble également équilibré en termes de phonèmes que nous utiliserons en phase de test. Ces ensembles sont identiques à ceux utilisés dans (Abderrazek *et al.*, 2023).
Le corpus **Common Phone** est un corpus équilibré en termes de genre, multilingue et aligné phonétiquement, dérivé du projet Common Voice de Mozilla. Seuls les enregistrements en français ont été utilisés dans ce travail. Nous avons également équilibré cet ensemble de données en termes de phonèmes et de genre pour assurer l'impartialité de nos modèles affinés du point de vue de ces derniers.
Le corpus **C2SI** comprend 87 patients traités pour un cancer buccal ou oropharyngé, ainsi que 41 contrôles sains (HC). Les patients et les HC ont été enregistrés à l'IUCT Oncopole, à Toulouse, France. Les patients ont été confrontés à plusieurs tâches : /a/ tenu, lecture de phrases, lecture de passages de texte (LEC), production de pseudo-mots (DAP) ainsi que diverses autres tâches prosodiques. À l'aide des enregistrements de certaines de ces tâches, des experts ont évalué la sévérité (degré d'altération du signal de parole) et l'intelligibilité de la production de parole des patients sur une échelle de 0 - forte altération - à 10 - discours parfait ; l'intelligibilité étant définie ici comme "la capacité d'un auditeur à reconnaître les mots et/ou les sons de la parole produite par le locuteur" (Astésano *et al.*, 2018). Les enregistrements des tâches LEC et DAP sont utilisés pour tester nos modèles. Ils seront désignés ici par C2SI-LEC et C2SI-DAP. Pour correspondre à la sélection effectuée dans un travail précédent, nous testerons nos modèles sur 24 HC (parmi les locuteurs à disposition, tous n'ayant pas

réalisé les deux tâches C2SI-LEC et C2SI-DAP), tous enregistrés dans les mêmes conditions. Nous confronterons nos résultats aux évaluations perceptives incluant 82 patients parmi les 87 – 5 patients n'ont pas été évalués par les experts.
Le tableau 1 détaille le nombre de trames alignées sur les phonèmes utilisés pour chaque corpus, ainsi que le nombre d'heures qu'elles représentent. Étant donné que certaines trames se chevauchent, ce nombre d'heures est supérieur à la somme des durées des enregistrements. Cependant, cette valeur reflète ce que le modèle voit en entrée. Tous les corpus mentionnés ci-dessus sont alignés phonétiquement avec 31 phonèmes et un silence. Ces 31 phonèmes comprennent quatre archi-phonèmes : /Ê/ = {e, ɛ}, /Û/ = {œ, ø}, /Ô/ = {o, ɔ}, et /µ/ = {œ̃, ɛ̃}. L'utilisation de ces archi-phonèmes permet de neutraliser les oppositions de voyelles moyennes.

TABLE 1 – Usage, nombre de trames et durée totale des données audios pour chaque corpus utilisé.

Corpus	Usage	#trames	#heures
BREF	entraînement, validation	3,118k	110h
Common Phone	entraînement, validation	236k	8.3h
BREF-Int	test	85k	3h
C2SI-LEC (HC)	test	43k	1.5h
C2SI-DAP (HC)	test	73k	2.5h

3 Modèles

Le modèle de réseau neuronal convolutif choisi a été précédemment entraîné dans (Abderrazek *et al.*, 2023) sur le corpus BREF décrit dans la section 2 . Il est composé de deux couches de convolution combinées avec des couches de pooling maximales. L'entrée du modèle correspond à une fenêtre glissante de 11 trames acoustiques de 20ms espacées de 10ms, où chaque trame correspond à un vecteur *Mel-Filterbanks* extrait du signal audio, ainsi qu'à ses dérivées première et seconde. Le modèle a ainsi une fenêtre de 120ms centrée sur le phonème à prédire. Une fois le CNN appliqué, la sortie est ensuite aplatie avant d'être donnée en entrée de trois couches denses, détaillées ci-dessous. La sortie de ce modèle est notre *baseline*.

Concernant la partie Wav2Vec2, les modèles de LeBenchmark2.0 (Parcollet *et al.*, 2023) sont utilisés. Comme nous allons appliquer ces modèles sur de la parole pathologique de locuteurs francophones dans la suite de nos travaux, nous avons ciblé les modèles Wav2Vec2 pré-entraînés exclusivement sur du français. En effet, combiner les phonèmes de plusieurs langues pourrait compliquer l'analyse entre les phonèmes sains et pathologiques. Les modèles LeBenchmark existent avec différentes architectures (6, 12, 24 ou 48 couches cachées) ainsi que des tailles de corpus de pré-entraînement différentes. Dans ce travail, nous comparerons les résultats obtenus avec des modèles contenant 6, 12 ou 24 couches cachées, respectivement appelés *light*, *base* et *large*, ainsi que pré-entraînés sur 3k ou 14k heures de parole française. Wav2Vec2 fonctionne avec des fenêtres de 25ms, espacées approximativement toutes les 20ms, ce qui implique un recouvrement de 5ms environ. Afin d'émuler la fenêtre de 120ms du CNN, nous donnons à Wav2Vec2 des fichiers audio correspondant à six fenêtres superposées de 25ms. Une telle architecture nous donne une fenêtre de 127ms environ, qui est très proche de la durée du contexte utilisé pour le CNN.

La sortie des modèles CNN ou Wav2Vec2, une fois aplatie, passe au travers de trois couches denses de dimension 1024, dédiées à la tâche de classification en phonèmes. La taille de la couche d'aplatissement dépend de la taille de la sortie de l'encodeur utilisé (CNN ou Wav2Vec2, et la taille

du modèle Wav2Vec2 choisi). Le phonème de sortie est ensuite sélectionné en appliquant un softmax sur les 32 valeurs de sortie.

Les corpus utilisés pour la phase d'entraînement ont été répartis de manière aléatoire en deux sous-ensembles équilibrés en termes de phonèmes : entraînement (90% des données) et validation (10% restants). Tous nos modèles Wav2Vec2 ont été affinés en utilisant l'outil SpeechBrain durant 15 époques (choix empirique). Le modèle présentant le meilleur taux d'erreur phonétique sur le jeu de validation a été choisi pour l'inférence sur les ensembles de données de test. Le classifieur utilise un optimiseur Adadelta avec un taux d'apprentissage initial de 0.9, pour améliorer une *cross-entropy loss* appliquée sur la classification de phonèmes. L'architecture de Wav2Vec2 - lorsqu'elle est affinée - utilise un optimiseur Adam avec un taux d'apprentissage initial de 1.10^{-4}. Les recettes SpeechBrain sont disponibles publiquement sur un répertoire GitHub [1].

4 Résultats expérimentaux

4.1 Comparaison des modèles Wav2Vec2

Des expériences ont été menées pour étudier l'impact des facteurs suivants : (1) l'affinage de Wav2Vec2, (2) les données de pré-entraînement, (3) la taille du modèle, et (4) les données d'affinage. Le tableau 2 résume les taux de bonne classification obtenus pour chaque modèle entraîné. Etant donné que la distribution des phonèmes n'est pas équilibrée dans les jeux C2SI, les taux ci-dessous sont équilibrés par phonème. Mathématiquement, il s'agit de la moyenne des taux de bonne classification obtenus pour chaque phonème. Ainsi, même lorsque ce n'est pas précisé, les taux de bonne classification sont systématiquement équilibrés.

Pour analyser l'impact de l'affinage de Wav2Vec2 (1), deux modèles *14k-large* ont été utilisés, seulement l'un des deux ayant été affiné. En comparant ainsi les modèles *14k-large Frozen* et *14k-large*, nous pouvons observer que l'affinage améliore les taux de bonne classification sur les trois jeux de test, de manière significative (les intervalles de confiance ne se recouvrent pas). Ainsi, affiner un modèle Wav2Vec2 sur des données du même type est bénéfique au niveau de cette métrique.

Ensuite, l'impact des données de pré-entraînement (2) a été mesuré en utilisant des modèles LeBenchmark ayant différents corpus de pré-entraînement. Nous nous sommes appuyés sur les familles de modèles 3k et 14k, qui sont entraînés respectivement sur 3 000 et 14 000 heures de parole française. Ils incluent de la parole française lue, jouée, spontanée et professionnelle, avec du français neutre ou comportant un accent. Ces deux ensembles de données varient fortement au niveau de la variété d'accents présents. En effet, l'ensemble de pré-entraînement 3k contient principalement des livres audio et des émissions de radios françaises, avec moins de 1% de parole comportant un accent africain dont nous avons connaissance - les livres audio ne fournissent pas d'informations sur les accents de leurs locuteurs. Malheureusement, nous ne pouvons pas être certains que les locuteurs des radios françaises soient des locuteurs francophones. Cependant, l'ensemble de pré-entraînement 14k contient environ 4700 heures de parole issue du Parlement Européen, qui présentent au minimum trois accents supplémentaires - belge, suisse et italien d'Aoste, ainsi que des quantités plus négligeables de radios africaines, qui incluent des accents maliens et nigériens. D'après nos résultats, nous pouvons observer que l'ajout d'une variabilité linguistique au travers de différents accents français, n'apporte pas d'amélioration significative des taux de bonne classification.

1. github.com/MaloMn/wav2vec2-phone-classification

TABLE 2 – Taux de bonne classification équilibrés (en %) obtenus sur chaque jeu de test. Les intervalles de confiances sont calculés avec une approche Bootstrap (Ferrer & Riera, 2024). Les taux obtenus pour le CNN ont été repris des travaux de (Abderrazek *et al.*, 2023), et ne présentaient pas d'intervalle de confiance.

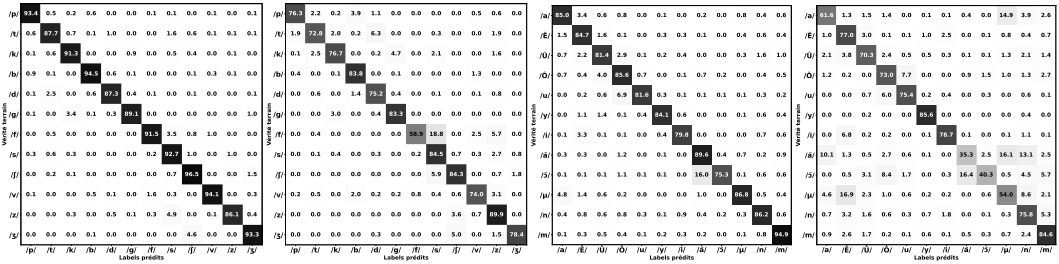
Modèle	Jeu(x) d'affinage	BREF-Int ↑	C2SI-LEC ↑ (locuteurs HC)	C2SI-DAP ↑ (locuteurs HC)
CNN, *Baseline*	BREF	81.4	72.2	69.2
14k-large Frozen	BREF	83.5±0.2	66.9±0.6	66.9±0.4
14k-large	BREF	87.6±0.2	70.2±0.6	70.6±0.4
14k-light	BREF	81.8±0.2	57.0±0.6	57.3±0.4
3k-large	BREF	**88.3±0.2**	70.6±0.6	71.3±0.4
3k-base	BREF	84.9±0.2	48.1±0.6	50.1±0.4
14k-large	BREF, CP	87.4±0.2	72.1±0.5	73.3±0.4
14k-light	BREF, CP	82.9±0.3	64.1±0.6	63.7±0.4
3k-large	BREF, CP	**88.3±0.2**	**72.6±0.6**	**73.9±0.4**
3k-base	BREF, CP	84.9±0.2	61.4±0.6	62.5±0.4

En ce qui concerne la taille du modèle, plusieurs tailles de modèles LeBenchmark (3) ont été testées : *light*, *base* et *large*, avec respectivement 6, 12 et 24 couches cachées. Nos résultats montrent qu'utiliser un modèle *large* (avec 24 couches cachées) offre de meilleurs taux de bonne classification que des modèles plus petits (6 et 12 couches cachées), et ce de manière significative. Malheureusement, comme LeBenchmark ne propose ni de modèles 14k-base et 3k-light, nous ne pouvons pas comparer les performances des modèles 3k-base et 14k-light. Néanmoins, nos résultats semblent montrer que les modèles plus petits ont un pouvoir de généralisation plus faible que les plus gros modèles sur nos jeux de données C2SI non vus des modèles.

Enfin, pour analyser l'impact des jeux d'affinage (4), nous ajoutons un jeu d'entraînement supplémentaire. Le travail antérieur utilisait exclusivement l'ensemble de données BREF pour entraîner l'architecture CNN. Tous les modèles mentionnés ci-dessus ont donc été ré-entraînés en combinant les corpus BREF et Common Phone. D'après les résultats obtenus, l'ajout de discours lu provenant d'autres corpus dans le processus d'affinage améliore la généralisation aux deux ensembles de données C2SI sur tous les modèles que nous avons affinés de manière significative. Les taux de bonne classification sont également similaires sur l'ensemble de données BREF-Int : les intervalles de confiance se chevauchent tous, sauf pour le modèle *14k-light*, qui est significativement meilleur une fois affiné sur la combinaison de BREF et Common Phone. Ainsi, il semble intéressant d'inclure d'autres ensembles de données à l'étape d'affinage, même s'ils ne représentent pas un pourcentage important des données d'entraînement.

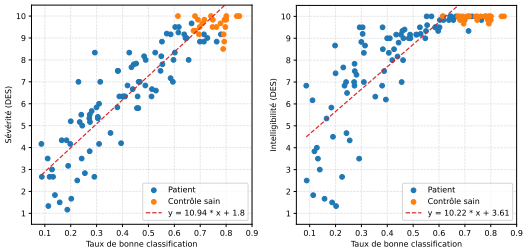
4.2 Wav2Vec2 et CNN

Le meilleur modèle Wav2Vec, par rapport aux résultats évoqués ci-dessus, nous permet d'améliorer de 6.9% les taux de bonne classification sur BREF-Int par rapport au CNN. Sur C2SI-LEC, la différence n'est pas significative. En revanche, on observe une différence significative sur C2SI-DAP, avec une amélioration de 4.7%. En comparant le nombre de paramètres des modèles que nous manipulons, il est intéressant de noter que le meilleur modèle Wav2Vec2 compte 330 millions de paramètres, contre 10 millions pour le CNN. Aussi, les modèles Wav2Vec2 plus petits *base* et *light* – ayant



(a) BREF-Int, obstruants (b) C2SI-LEC, obstruants (c) BREF-Int, vocaliques (d) C2SI-LEC, vocaliques

FIGURE 1 – Matrices de confusion obtenues sur les jeux BREF-Int – figures 1a et 1c – et C2SI-LEC – figures 1b et 1d – (locuteurs HC uniquement).



(a) Sévérité (DES), corrélation r = 0.90 (b) Intelligibilité (DES), corrélation r = 0.80

FIGURE 2 – Nuages de points des patients et HC C2SI suivant les niveaux associés aux évaluations perceptives des experts et les taux de bonne classification équilibrés obtenus.

respectivement 90 et 26 millions de paramètres – ne parviennent pas à généraliser aussi bien que le CNN sur les jeux C2SI. Cet écart est important, mais ces deux types de modèles ne présentent pas non plus la même architecture, et une comparaison basée uniquement sur le nombre de paramètres serait biaisée. Néanmoins, cette différence de taille implique que l'empreinte carbone de nos inférences sera nécessairement plus élevée lors de l'utilisation de modèles Wav2Vec2. A propos des jeux de données utilisés, le meilleur modèle a été entraîné sur davantage de données que le CNN. Nos résultats montrent aussi qu'utiliser uniquement BREF lors de l'affinage ne permet pas de généraliser aussi bien que le CNN sur les jeux C2SI.

Pour nous assurer que nos modèles ne sur-ajustent pas certains phonèmes, et que les confusions restent explicables, des matrices de confusion ont été générées sur BREF-Int et C2SI-LEC, en utilisant le modèle *3k-large* affiné sur BREF et Common Phone. Les figures 1a à 1d présentent des parties spécifiques de ces matrices, dédiées aux phonèmes obstruants et vocaliques. Le choix de restreindre notre analyse à ces phonèmes nous permet de réaliser une comparaison directe avec l'analyse effectuée dans (Abderrazek *et al.*, 2023). La comparaison entre les résultats obtenus et ceux obtenus précédemment montre que Wav2Vec2 diminue dans la plupart des cas les confusions observées. Sur les phonèmes obstruants, /ʒ/ était confondu avec /ʃ/ dans 9% des cas sur BREF-Int sur le précédent travail, contre 4.6% maintenant. La confusion entre ces deux phonèmes est donc toujours présente, mais elle survient dans la moitié des cas par rapport à précédemment. Cette diminution se retrouve aussi sur le jeu C2SI : alors que /p/ était précédemment confondu avec /t/ dans 9.3% des cas, ce n'est le cas que 2.2% du temps avec notre architecture. Nous retrouvons également ici

les mêmes causes pour les confusions importantes : soit la perte du lieu d'articulation (caractère aigu – /f/ → /s/, caractère compact – /ʃ/ → /s/), soit la confusion du voisement – /t/ → /d/. Sur les phonèmes vocaliques, nous retrouvons également de fortes confusions liées aux voyelles orales et aux consonnes nasales sur le jeu C2SI-LEC. Là où /ã/ et /a/ étaient confondus dans 11.5% des cas, ils le sont toujours dans 10.1% des cas. /ã/ est aussi davantage confondu maintenant avec /n/ (13.1%, contre 8.4% précédemment), et est moins confondu avec /m/ (2.5%, contre 6.2% précédemment). Ces confusions, expliquées précédemment par le changement d'accents des locuteurs (Parisien.ne.s pour BREF-Int, et Toulousain.e.s pour C2SI) se retrouvent ici, avec l'utilisation d'un autre modèle. Nos résultats viennent donc appuyer les résultats précédemment obtenus, et montrent que les données utilisées pour l'affinage et les différences de domaine entre jeux de données restent une problématique non négligeable pour ce type d'architecture. Sur BREF-Int, nous retrouvons également une confusion forte entre /ã/ et /c̃/, avec une confusion mutuelle de 16.4%, mais moins intense qu'en utilisant un CNN – 20.6%. Ces résultats sont importants car ils montrent que notre modèle est aussi sensible aux prononciations atypiques (ici, un accent régional), ce qui est souhaitable lors de l'analyse de la parole pathologique.

4.3 Application à la parole pathologique

La robustesse de notre modèle, ainsi que sa capacité à généraliser à d'autres jeux de données ayant été montrées, nous allons observer si les taux de bonne classification peuvent corréler positivement avec les évaluations des 6 experts de C2SI. Les jugements des experts, en termes de scores de sévérité et d'intelligibilité sur la tâche de description d'image (DES) ont été moyennés, puis comparés aux taux de bonne classification. Les figures 2a et 2b comparent sous la forme de nuages de points les scores perceptifs et les taux de bonne classification associés aux enregistrements de parole des patients et HC du corpus C2SI. Les courbes de régression linéaire ont également été tracées. Les coefficients de corrélation de Pearson élevés – 0.90 avec la sévérité et 0.80 avec l'intelligibilité – confirment que les mesures perceptives peuvent être estimées à l'aide des taux de bonne classification équilibrée des phonèmes de notre modèle *3k-large* affiné sur BREF et Common Phone. Ces valeurs sont semblables à celles obtenues précédemment avec le CNN (Abderrazek *et al.*, 2023) dont les taux de bonne classification en phonèmes corrélaient à 0.91 et 0.81 avec la sévérité et l'intelligibilité respectivement. Ces derniers résultats confirment qu'une représentation de la parole basée sur un modèle de type Wav2Vec2 convient bien à une analyse phonétique de la parole pathologique.

5 Conclusion

Dans ce travail, nous avons montré qu'un modèle basé sur Wav2Vec2 surpasse un CNN dans la classification des phonèmes, tout en préservant certaines spécificités linguistiques, telles qu'un accent régional. Nos résultats valident non seulement l'efficacité de l'approche basée sur Wav2Vec2, mais soulignent également l'importance de prendre en compte l'architecture du modèle et la diversité des données d'entraînement pour des performances optimales. Les travaux futurs incluent l'application du concept de NCD développé dans (Abderrazek *et al.*, 2023) pour analyser les couches cachées de notre architecture affinée, et analyser comment ce nouveau modèle influence la détection des traits phonétiques, qui sont cruciaux pour expliquer les phonèmes prédits. À son tour, cette interprétabilité est nécessaire pour avoir une analyse objective de la parole d'un patient, ce qui améliorerait les techniques de rééducation mises en place par les cliniciens.

Références

ABDERRAZEK S., FREDOUILLE C., GHIO A., LALAIN M., MEUNIER C. & WOISARD V. (2023). Interpreting deep representations of phonetic features via neuro-based concept detector : Application to speech disorders due to head and neck cancer. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, **31**, 200–214. DOI : 10.1109/TASLP.2022.3221039.

ASTÉSANO C., BALAGUER M., FARINAS J., FREDOUILLE C., GAILLARD P., GHIO A., LAARIDH I., LALAIN M., LEPAGE B., MAUCLAIR J., NOCAUDIE O., PINQUIER J., PONT O., POUCHOULIN G., PUECH M., ROBERT D., SICARD E. & WOISARD V. (2018). Carcinologic speech severity index project : A database of speech disorder productions to assess quality of life related to speech after cancer. In N. CALZOLARI, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, K. HASIDA, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK, S. PIPERIDIS & T. TOKUNAGA, Éds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan : European Language Resources Association (ELRA).

DARLEY F. L., ARONSON A. E. & BROWN J. R. (1969). Clusters of deviant speech dimensions in the dysarthrias. *Journal of Speech and Hearing Research*, **12**(3), 462–496. DOI : 10.1044/jshr.1203.462.

FAVARO A., TSAI Y.-T., BUTALA A., THEBAUD T., VILLALBA J., DEHAK N. & MORO-VELÁZQUEZ L. (2023). Interpretable speech features vs. dnn embeddings : What to use in the automatic assessment of parkinson's disease in multi-lingual scenarios. *Computers in Biology and Medicine*, **166**, 107559. DOI : https ://doi.org/10.1016/j.compbiomed.2023.107559.

FERRER L. & RIERA P. (2024). Confidence Intervals for evaluation in machine learning.

HERNANDEZ A., PÉREZ-TORO P. A., NOETH E., OROZCO-ARROYAVE J. R., MAIER A. & YANG S. H. (2022). Cross-lingual Self-Supervised Speech Representations for Improved Dysarthric Speech Recognition. In *Proc. Interspeech 2022*, p. 51–55. DOI : 10.21437/Interspeech.2022-10674.

JAVANMARDI F., KADIRI S. R. & ALKU P. (2024). Pre-trained models for detection and severity level classification of dysarthria from speech. *Speech Communication*, p. 103047. DOI : https ://doi.org/10.1016/j.specom.2024.103047.

KLUMPP P., ARIAS T., PÉREZ-TORO P. A., NOETH E. & OROZCO-ARROYAVE J. (2022). Common phone : A multilingual dataset for robust acoustic modelling. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 763–768, Marseille, France : European Language Resources Association.

LAMEL L. F., GAUVAIN J.-L., ESKÉNAZI M. *et al.* (1991). Bref, a large vocabulary spoken corpus for french. *Eurospeech'91, Italy*, **22**(28), 50.

LUNDBERG S. M. & LEE S.-I. (2017). A unified approach to interpreting model predictions. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30 : Curran Associates, Inc.

PARCOLLET T., NGUYEN H., EVAIN S., BOITO M. Z., PUPIER A., MDHAFFAR S., LE H., ALISAMIR S., TOMASHENKO N., DINARELLI M., ZHANG S., ALLAUZEN A., COAVOUX M., ESTEVE Y., ROUVIER M., GOULIAN J., LECOUTEUX B., PORTET F., ROSSATO S., RINGEVAL F., SCHWAB D. & BESACIER L. (2023). Lebenchmark 2.0 : a standardized, replicable and enhanced framework for self-supervised representations of french speech.

TOM DIECK T., PÉREZ-TORO P. A., ARIAS T., NOETH E. & KLUMPP P. (2022). Wav2vec behind the Scenes : How end2end Models learn Phonetics. In *Proc. Interspeech 2022*, p. 5130–5134. DOI : 10.21437/Interspeech.2022-10865.

TU M., BERISHA V. & LISS J. (2017). Interpretable Objective Assessment of Dysarthric Speech Based on Deep Neural Networks. In *Proc. Interspeech 2017*, p. 1849–1853. DOI : 10.21437/Interspeech.2017-1222.

XU L., LISS J. & BERISHA V. (2023). Dysarthria detection based on a deep learning model with a clinically-interpretable layer. *JASA Express Letters*, **3**(1), 015201. DOI : 10.1121/10.0016833.

YEO E. J., CHOI K., KIM S. & CHUNG M. (2023a). Automatic severity classification of dysarthric speech by using self-supervised model with multi-task learning. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 1–5. DOI : 10.1109/ICASSP49357.2023.10094605.

YEO E. J., CHOI K., KIM S. & CHUNG M. (2023b). Speech intelligibility assessment of dysarthric speech by using goodness of pronunciation with uncertainty quantification. In *Proc. INTERSPEECH 2023*, p. 166–170. DOI : 10.21437/Interspeech.2023-173.