Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

Comparaison de mesures pour la détection automatique de déviance dans la dysarthrie ataxique

Natacha Miniconi[1] Cédric Gendrot[1] Angélina Bourbon[1] Leonardo Lancia[2]
Cécile Fougeron[1]
(1) Laboratoire de phonétique et phonologie, UMR 7018 / Université Sorbonne Nouvelle
prenom.nom@sorbonne-nouvelle.fr
(2) Laboratoire Parole et Langage/ (CNRS AMU)
lancia.leonardo@univ-amu.fr

RÉSUMÉ

Cette étude explore l'utilisation d'un Réseau de Neurones Convolutifs (CNN) pour distinguer la parole de patients dysarthriques ataxiques de celle de locuteurs neurotypiques, en utilisant diverses entrées. L'objectif est d'extraire automatiquement des informations pertinentes sur les troubles de la parole. Le CNN est utilisé pour exploiter les caractéristiques temporelles et spectrales des signaux de parole via des spectrogrammes, des trajectoires de formants et des courbes de modulation cepstrale. Comparé à un Multi-Layer Perceptron (MLP) alimenté par des mesures acoustico-phonétiques ciblées sur la modulation cepstrale, le CNN présente de meilleurs scores de classification dans la distinction entre dysarthrie et non dysarthrie, en particulier avec la modulation cepstrale. La population CTRL obtient de meilleurs taux de classification que la population SCA avec un MLP, alors qu'on on observe l'inverse avec un CNN.

ABSTRACT

Comparison of measures for automatic detection of deviance in ataxic dysarthria.

This study explores the use of a Convolutional Neural Network (CNN) to distinguish the speech of ataxic dysarthric patients from that of neurotypical speakers, using various inputs. The aim is to automatically extract relevant information about speech disorders. The CNN is used to exploit the temporal and spectral characteristics of speech signals via spectrograms, formant trajectories and cepstral modulation curves. Compared with a Multi-Layer Perceptron (MLP) fed with targeted acoustic-phonetic measurements on cepstral modulation, the CNN shows better classification scores in the distinction between dysarthria and non-dysarthria, particularly with cepstral modulation. The CTRL population obtained better classification rates than the SCA population with an MLP, while the opposite was observed with a CNN.

MOTS-CLÉS : dysarthrie, deep learning, detection de déviance, acoustique.

KEYWORDS: dysarthria, deep learning, deviance detection, acoustic.

1 Introduction

L'analyse et la compréhension des signaux de parole permettent d'extraire de multiples informations sur le locuteur. Cela rejoint l'un des principaux buts de la phonétique clinique qui est d'affiner la caractérisation de la parole des patients en identifiant avec précision les aspects déviants. Les troubles de l'articulation, par exemple, sont prépondérants et peuvent être caractérisés par des difficultés à

Actes des 35èmes Journées d'Études sur la Parole, pages 281–290.

atteindre les cibles articulatoires dans le temps et l'espace.

Dans cette étude, nous nous intéresserons à une parole pathologique particulière, celle des patients atteints d'ataxies spinocérébelleuses (SCA), un groupe hétérogène de maladies neurodégénératives dûes à une atteinte du cervelet et pouvant présenter des symptômes sur l'ensemble de la sphère motrice. Parmi ces symptômes, nous pouvons retrouver une dysarthrie qualifiée d'ataxique (Darley *et al.*, 1969). Elle se caractérise, entre autres, par des altérations de l'articulation des consonnes et voyelles et des allongements des durées segmentales (Shalling *et al.*, 2007; Brendel *et al.*, 2015; Schmitz-Hübsch *et al.*, 2011).

Généralement, l'évaluation clinique des troubles de la parole est faite à l'oreille et quantifiée à l'aide de scores basés sur des relevés d'erreurs de prononciation (Laganaro *et al.*, 2020). Cette caractérisation peut aussi être réalisée par des mesures acoustiques temporelles ou spectrales quantifiables, comme le débit de parole ou des mesures de formants sur les voyelles (Brendel *et al.*, 2015; Audibert & Fougeron, 2012). Ces dernières mesures nécessitent souvent une segmentation et/ou une intervention manuelle sur le signal de parole qui requiert une expertise approfondie et un investissement temporel conséquent. À ce titre, il est possible d'appliquer des mesures plus globales sur des séquences de parole non segmentées en phonèmes ou syllabes ; cette approche est déjà adoptée par plusieurs études sur la parole pathologique avec l'utilisation de modulation cepstrale (Slis *et al.*, 2021) ou de paramètres opensmile (Kodrasi *et al.*, 2021), de mesures acoustico-phonétiques ciblées prises sur les formants (Wang *et al.*, 2016) afin de discriminer la parole pathologique. D'autres études ont utilisé des segmentations à l'aide d'aligneur automatiques permettant de distinguer dysarthrique et non dysarthrique au niveau phonémique (Laaridh *et al.*, 2016). En contraste de ces mesures interprétables, des mesures non interprétables sont également utilisées pour capturer l'articulation comme X-vector (Favaro *et al.*, 2023).

L'utilisation des CNN pour les recherches phonétiques a introduit d'autres types d'informations pertinentes en appliquant des mesures non interprétables pour la caractérisation de l'articulation en y introduisant en entrée des images de spectrogrammes (Faragó *et al.*, 2022; Kim & Gendrot, 2022), mais aussi, des caractéristiques d'énergie de la banque de filtres Mel affectée en entrée au CNN (Abderrazek *et al.*, 2020). La présente étude cherche à évaluer la pertinence de ces différentes informations extraites d'un enregistrement audio avec un minimum d'intervention manuelle pour caractériser les troubles articulatoires dans la dysarthrie.

Ainsi, la question de recherche centrale se formule comme suit : quelles informations spécifiques s'avèrent pertinentes pour distinguer la parole de patients dysarthriques de celle de locuteurs non dysarthriques ?

Pour répondre à cette interrogation, nous proposons d'utiliser un Réseau de Neurones Convolutifs (CNN) ayant comme tâche une classification binaire : dysarthrique ou non-dysarthrique avec plusieurs types d'entrées tels que :

— Images de spectrogrammes
— Images capturant des trajectoires des formants
— Images capturant la modulation cepstrale des productions

Afin d'évaluer la pertinence des classifications produites par le CNN, nous confrontons ses résultats à une classification basée sur des mesures issues d'une expertise phonétique. Pour ce faire, nous utiliserons un perceptron multicouche (MLP) configuré pour utiliser des mesures ciblées sur certains événements de la modulation cepstrale (par exemple les maxima). Ce type de modèle MLP a déjà été utilisé sur d'autres types de dysarthrie en l'alimentant de mesures phonétiques (Alshammri

et al., 2023).

2 Méthodologie des entrées du CNN et du MLP

Les enregistrements utilisés pour cette étude proviennent des projets SpeechN'Co (n° ID-RCB : 2019-A02553-54) et ChaSpeePro (CRSII5_202228). Ils incluent 30 locuteurs, 16 hommes et 14 femmes (*moy.* = 53.4, 23<>72 ans), tous porteurs d'ataxies spinocérébelleuses. Ils présentent tous une dysarthrie, avec une sévérité variable évaluée à l'aide du score perceptif de la batterie d'évaluation BECD (Bourbon *et al.*, 2023). Nous avons comparé ces enregistrements des locuteurs dysarthriques à ceux de deux groupes contrôles, chacun constitués de 30 locuteurs qui ont été tirés aléatoirement à partir des deux bases de données, comprenant des locuteurs neurotypiques âgés de 24 à 90 ans. Le groupe CTRL1 comprend 14 hommes et 16 femmes (*moy.* = 60.38 ans, 24<>90 ans) et le groupe CTRL2 comprend 11 hommes et 19 femmes (*moy.* = 57.7 ans, 25<>82 ans) Le matériel linguistique enregistré est extrait du protocole MonPaGe-MoSpeeDi et consiste en trois séquences glides-voyelles ayant un sens en français. Ces séquences comportent chacune trois syllabes : 'aille-aille-aille' /ajajaj/, 'ouille-ouille-ouille' /ujujuj/ et 'oui-oui-oui-' /wiwiwi/. Chaque locuteur avait pour consigne de produire ces séquences de façon continue, sans pause entre les syllabes, à un débit et une intensité confortable. Ce matériel a été construit pour évaluer des modulations articulatoires sur la suite de trois syllabes via des modulations acoustiques continues sur le signal acoustique produit (Lévêque *et al.*, 2022). La Table 1 présente le nombre d'enregistrements utilisés dans l'étude.

Logatomes	SCA	CTRL1	CTRL2
ajajaj	117	119	120
ujujuj	112	119	119
wiwiwi	112	119	120
Total	341	357	359

TABLE 1 – Répartition du nombre de fichiers audios pour les groupes de locuteurs avec parole dysarthrique(SCA) et sans parole dysarthrique(CTRL1, 2) correspondant chacun à une séquence que le locuteur doit produire.

2.1 Types d'informations acoustiques extraites du signal

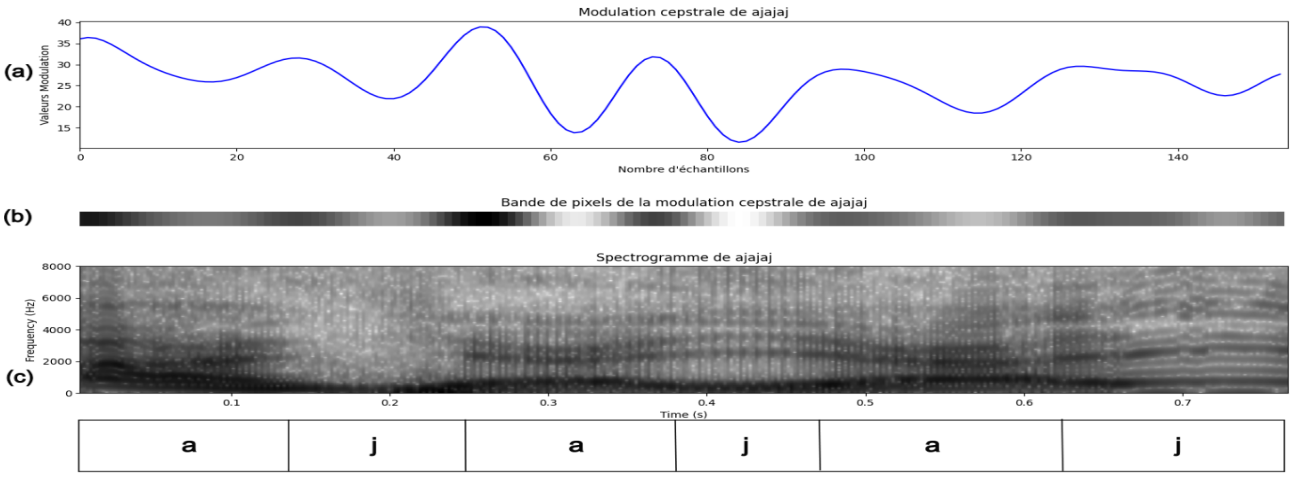


FIGURE 1 – Représentation des différents types de mesures utilisées dans le CNN pour le locuteur : SCA_F_AB03. La modulation cepstrale notée (a) devient (b) pour l'entrée du CNN. Les trajectoires formantiques subissent la même transformation. Les spectrogrammes ont été extraits sur la séquence entière (c) en tant qu'entrée pour le réseau.

Noms entrées	Types d'informations acoustiques extraites du signal
Mod_Cepstr	La modulation cepstrale permet de quantifier le degré de changement global des caractéristiques spectrales (Goldstein, 2019). Elle est ici sous la forme d'une courbe où les pics indiquent des changements d'état avec une forte différence d'énergie spectrale, tandis que les creux reflètent des périodes de stabilité avec une faible différence. Afin d'obtenir les modulations cepstrales, suivant la procédure élaborée par Leonardo Lancia (Slis *et al.*, 2021), les coefficients cepstraux en Mel (MFCC) ont été calculés grâce à la bibliothèque *Librosa* entre 300Hz et 8KHz, avec la possibilité de personnaliser des paramètres tels que la longueur de la fenêtre d'analyse (définie à 25 ms), le pas temporel (fixé à 5 ms) et le nombre de MFCC à extraire (choisi à 13). Suite à l'extraction des MFCCs, à chaque pas de l'analyse, la valeur absolue de la différence entre les valeurs successives de chaque coefficient a été calculée. Ensuite, chaque série chronologique obtenue a été soumise à un filtre passe-bas avec une fréquence de coupure fixée à 12 Hz. Afin de pouvoir utiliser les informations des modulations cepstrales, celles-ci ont par la suite subi une transformation visuelle en représentant chaque courbe sous la forme d'une bande de pixels. Dans cette représentation, l'intensité de la courbe est reflétée par la teinte des pixels, où une teinte plus foncée correspond à une intensité plus élevée de la courbe. L'échelle a été normalisée sur la totalité du corpus. Sa représentation est visible sur la Figure 1b.
Spectro	Les spectrogrammes, capturant les déformations potentielles des voyelles et des consonnes induites par la maladie (Shalling *et al.*, 2007). Les fréquences en Hz retenues pour les spectrogrammes se situent dans la plage de 0 à 8 000 Hz.
F1, F2, F3	Les spectrogrammes sont susceptibles de comporter beaucoup de bruit. De ce fait, nous allons également extraire les trajectoires des formants F1, F2, F3 sur la durée totale de la séquence produite grâce à la fonction *To Formant (burg)* de Praat qui extrait la valeur de chaque formant toutes les 5ms. Après un lissage sur python, pour chaque formant, la trajectoire a été transformée en une bande de pixels, suivant le même processus appliqué à la courbe de modulation cepstrale. Les bandes de formant ont été empilées, cela permettant d'avoir une image comportant les trois bandes de formants (F1+F2+F3), comme le spectrogramme illustré figure 1c.
Mean pics	Moyennes des pics sur la courbe de modulation cepstrale correspondant au moment dans la production avec le plus de changement cepstraux (=aux transitions entre les phonèmes). Voir Figure 2.
Meanch	Moyenne de toutes les valeurs de la courbe de modulation cepstrale sur la production. Voir Figure 2.
SD Meanch	Ecart-type de toutes les valeurs de la courbe de modulation cepstrale sur la production.
EventDUR	La moyenne des durées entre deux minimums consécutifs. Voir Figure 2.
SD EventDUR	Ecart-type des durées entre les pics qui reflète la régularité des durées entre les segments.

TABLE 2 – Descriptions des différents types d'informations acoustiques extraites du signal

Les mesures ciblées (Mean pics, Meanch, SD Meanch, EventDUR, SD EventDUR) réalisées sur la modulation cepstrale développées dans des études ultérieures (Slis *et al.*, 2021; Lévêque *et al.*, 2022) sont basées sur une quantification de la modulation cepstrale durant la séquence produite, de façon à approximer les changements dans le temps du conduit vocal lors de la production. Les mesures ont permis de mettre en évidence des différences dans les schémas articulatoires des personnes atteintes de dysarthrie. Le même type de méthode a déjà été réalisé dans des études antérieures où ont été utilisées comme indice articulatoire les transitions : consonnes, voyelles (Mathad *et al.*, 2022; Xu *et al.*, 2022). Ci-dessous la figure 2 représentant la prise des mesures ciblées.

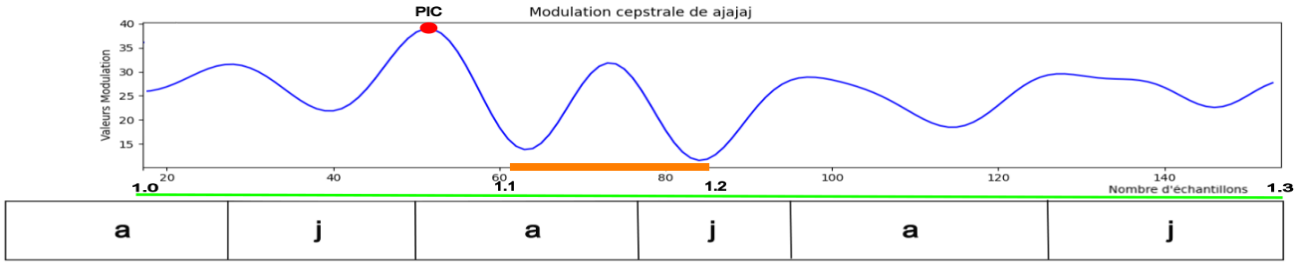


FIGURE 2 – Modulation cepstrale d'un *ajajaj* du locteur SCA_F_AB03. La valeur moyenne des 5 pics maximums représente la mesure Mean pics. EventDUR représente la durée moyenne de chaque intervalle de temps entre deux points de minimum consécutifs dans la modulation cepstrale. Une de ces durées est représentée entre le point 1.1 et 1.2 surlignée en orange. Meanch représente la valeur moyenne de toutes les valeurs sur la durée entre 1.0 et 1.3 surlignée en vert.

2.2 Élaboration du CNN et du MLP

Les CNN prennent comme entrées : Spectro, Mod_Cepstr et les Trajectoires F1, F2, F3 sous forme d'image. Le MLP quant à lui reçoit comme entrée des données structurées sous forme numérique,

comprenant des mesures spécifiques affectées sur la modulation cepstrale décrite dans la Table 2 (Mean pics, Meanch, SD Meanch, EventDUR, SD EventDUR). Ces mesures, préalablement calculées et organisées dans un tableau, ont servi d'entrée pour l'entraînement et pour le test du MLP.

Pour les deux modèles, la tâche de classification est identique : il s'agit de déterminer si les enregistrements vocaux proviennent d'un locuteur avec une parole dysarthrique ou non. Pour y parvenir, nous avons utilisé un corpus de 30 locuteurs ayant une parole dysarthrique (SCA) et de 30 locuteurs n'ayant pas de parole dysarthrique (CTRL1) constituant le premier essai appelé test1. Ensuite, pour valider notre méthode, nous avons effectué la même tâche de classification en utilisant cette fois les 30 locuteurs ayant une parole dysarthrique et 30 autres locuteurs n'ayant pas de parole dysarthrique issus groupe CTRL2. Cette classification constitue notre deuxième essai appelé test2. Ce test2 permet d'évaluer la stabilité des modèles en regardant si les résultats sont les mêmes avec un autre groupe CTRL.

Lors de la réalisation des deux modèles, la procédure "*Leave One Out*" (LOO) a été employée. Le modèle a été exécuté sur 60 itérations pour les corpus comprenant SCA et CTRL1 et CTRL2, en utilisant chaque locuteur comme locuteur de test une fois, tout en utilisant toutes les données des autres locuteurs pour l'entraînement. Cela nous permet d'éviter du biais ainsi que des bonnes performances dues au hasard ou bien à un choix particulier sur le corpus test (Berrar, 2019).

Pour chaque sortie de ces modèles, nous obtenons plusieurs scores de performance : l'*accuracy* de chaque locuteur, le *F-score* et le *rappel*. Ces deux derniers scores n'étant pas pertinents pour cette étude, ils ne seront pas mentionnés. L'*accuracy* correspond au nombre d'enregistrements correctement prédits sur l'ensemble d'enregistrement du locuteur.

L'architecture du CNN débute par une couche de convolution avec 64 filtres de taille 3×3 et un padding de type valid qui produit une sortie de taille réduite par rapport à l'entrée, suivi d'une normalisation par lots. Elle incorpore ensuite un *ZeroPadding2D* pour gérer la variation de la longueur des enregistrements audios. Chaque locuteur a fourni des enregistrements audios de durées différentes, ce qui a conduit à des variations dans la taille des images correspondantes aux différentes mesures. La couche de *ZeroPadding2D* a été employée pour pallier ce problème. Son rôle est d'ajuster la taille des images spectrogrammes plus courtes en ajoutant des "pixels de remplissage" autour d'elles. Ensuite, un *MaxPooling2D* a été ajouté avec un pool de taille 2×2, et une autre normalisation par lots. La seconde couche de convolution utilise 16 filtres de 3×3 avec padding de type valid, suivi d'une nouvelle normalisation. Après avoir aplati les données, le modèle applique un Dropout de 0.5, puis trois couches denses avec respectivement 16 et 8 unités avec activation ReLU, et enfin une unité avec activation sigmoïdale.

Quant au MLP, il a été utilisé la bibliothèque *sklearn*. La fonction d'activation des couches cachées est ReLU. La taille des couches cachée est de 100 neurones, nous avons choisi le même nombre de couches cachées que pour le CNN. D'autres paramètres, tels que le taux d'apprentissage constant, le nombre maximal d'itérations, l'optimiseur "adam" sont fixés aux valeurs par défaut. Une analyse en composantes principales (ACP) a été faite sur les mesures numériques données au MLP et a révélé que la première composante principale explique 50% de la variance, tandis que la deuxième en explique 30%. Ensemble, ces deux composantes couvrent donc 80% de la variance totale. Pour atteindre le seuil souhaité de 95% de la variance expliquée, nous nous retrouvons dans une situation où le nombre de composantes nécessaires dépasse celui des mesures originales, rendant l'approche peu avantageuse. L'apprentissage du modèle a donc été effectué sans réduction de dimension.

3 Résultats

Entrées CNN	SCA		CTRL1	
	Accuracy	SD Accuracy/min/max	Accuracy	SD Accuracy/min/max
Mod_Cepstr	**84%**	17% / 50% / 100%	**72%**	23% / 1% / 100%
Spectro	74%	30% / 1% / 100%	47%	37% / 0% / 100%
F1 + F2 + F3	72%	26% / 1 % / 100 %	42%	30% / 0% / 92 %
F1	73%	22% / 16% / 100%	38%	22% / 0% / 90%
F2	71%	23% / 17% / 100%	42%	20% / 1% / 83%
F3	72%	17% / 36% / 100%	30%	18% / 0% / 83%
Entrées MLP				
Mean pics + Meanch + SD Meanch + EventDUR + SD EventDUR	73%	30% / 1% / 100%	76%	30% / 30% / 100%
Mean pics	65%	31% / 1% / 100%	78%	20% / 33% / 100%
Meanch + SD Meanch	**74%**	27% / 0% / 100 %	**81%**	26% / 1% / 100%
EventDUR + SD EventDUR	70%	27% / 0 % / 100 %	71%	30% / 1% / 100 %

TABLE 3 – Performances du CNN et MLP lors du test 1 en fonction des différentes entrées et des groupes (SCA et CTRL1). Les performances sont estimées à partir du pourcentage de fichiers bien classés (*accuracy*), de la variabilité des *accuracy* entre les locuteurs en termes d'écart-type, minimum et maximum d'*accuracy* (SD/min/max *accuracy*).

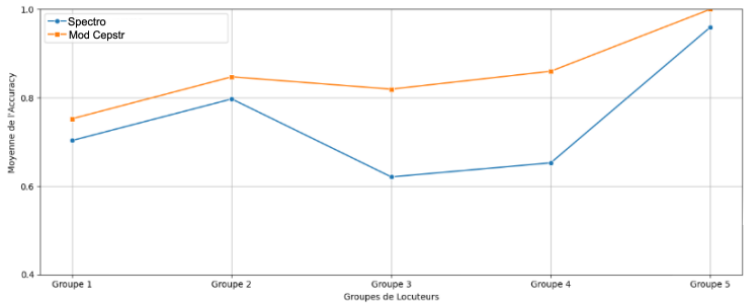
3.1 Effet du type d'entrée sur les performances de classification (CNN et MLP)

La Table 3 présente la moyenne des résultats de l'évaluation des performances du CNN et MLP avec chacune des entrées pour le groupe SCA et le groupe CTRL1 lors du test1. Les entrées provenant d'informations sur la modulation cepstrale montrent une meilleure performance face à celles provenant d'un spectrogramme. Les résultats montrent que la meilleure entrée du CNN est Mod_Cepstr avec 84% des enregistrements bien classés pour les locuteurs SCA et 72% pour les CTRL1. Lors du test2, avec la classification sur les mêmes locuteurs du groupe SCA et un groupe d'autres locuteurs contrôles (CTRL2), les résultats sont similaires. L'entrée Mod_Cepstr classifie correctement 80% des productions des locuteurs SCA et 71% des productions des locuteurs du groupe CTRL2. Les entrées du MLP extraites de la modulation cepstrale montrent également de bonnes performances en comparaison de celles issues d'un spectrogramme (Spectro, F1, F2, F3). La meilleure entrée du MLP est celle composée de Meanch et SD Meanch avec une *accuracy* moyenne de 74% pour les locuteurs SCA et 81% pour les locuteurs CTRL1. Pour la comparaison avec les locuteurs CTRL2, nous avons à nouveau peu de différences : une *accuracy* de 72 % pour les locuteurs SCA et 77 % pour les locuteurs CTRL2. Ce faible écart entre les groupes CTRL1 et CTRL2 s'observe par ailleurs sur la totalité des entrées. L'*accuracy* de l'entrée Spectro avec 74% pour les locuteurs SCA et 47% pour les locuteurs CTRL1 est légèrement inférieure à celle de l'entrée Mod_Cepstr pour les locuteurs SCA mais pour les locuteurs CTRL1, une chute des performances est observée où un enregistrement sur deux est mal classé. Quant aux entrées F1, F2, F3, elles sont inférieures ou égales à l'entrée Spectro, nous observons cette même chute de performances pour les groupes CTRL1 et 2, en particulier avec F3.

3.2 Performance de classification en fonction des locuteurs (CNN et MLP)

Un calcul du coefficient de corrélation de Spearman est réalisé afin de savoir si la sévérité des SCA est corrélée au taux d'*accuracy* pour les différentes entrées. Nous pouvons voir que les performances ne sont pas similaires en fonction de la sévérité des troubles de la parole : sans surprise, les locuteurs les plus sévères sont les mieux reconnus comme étant dysarthriques, à hauteur de 100% des enregistrements pour les locuteurs les plus sévères sur la base de l'entrée Mod_Cepstr. Pour les locuteurs les moins sévères, une baisse de performance à hauteur de 18% est observée. Une corrélation positive modérée significative est constatée entre l'*accuracy* et la sévérité pour l'entrée Mod_Cepstr ($\rho = 0.42$, $p < 0.05$), mettant en exergue l'influence de la sévérité. En revanche,

pour toutes autres entrées du CNN, la corrélation est quasi-nulle avec une p-value non significative ($\rho = -0.2$, $p > 0.05$). Le taux d'*accuracy* des performances de classification avec les entrées : Spectro, F1, F2, F3 n'est pas sensible à la sévérité de la dysarthrie. Sur la Figure 3 nous pouvons remarquer que Mod_Cepstr classe beaucoup mieux les plus sévères avec un faible creux dans les sévérités intermédiaires, à contrario de l'entrée Spectro qui n'est effectivement pas affecté par la sévérité.



Groupes	Sévérité (S_1)
Groupe 1	S_1 compris entre 1.4 et 3.4
Groupe 2	S_2 compris entre 4.6 et 6.5
Groupe 3	S_3 compris entre 7 et 8.8
Groupe 4	S_4 compris entre 9.2 et 12.80
Groupe 5	S_5 compris entre 13.4 et 17

FIGURE 3 – Graphique de l'accuracy des 5 groupes SCA dans l'ordre croissant de sévérité en fonction de l'accuracy. Les sévérités des locuteurs de chaque groupe sont indiquées dans le tableau à droite.

Nous retrouvons cette corrélation positive et significative entre les performances de classification du MLP et la sévérité de locuteurs SCA. Pour l'entrée détenant l'intégralité des mesures, le coefficient de corrélation indique une forte corrélation positive entre le taux d'*accuracy* et la sévérité ($\rho = 0.72$, $p < 0.001$). Enfin, l'entrée incluant EventDUR et SD EventDUR présente une corrélation positive et légèrement moins élevée que celle de l'entrée détenant l'intégralité des mesures ($\rho = 0.65$, $p < 0.001$). Nous obtenons les mêmes valeurs pour Mean pics. L'entrée Meanch + SD Meanch montre une corrélation positive légèrement moins élevée ($\rho = 0.58$, $p < 0.001$). L'apport du CNN par rapport au MLP est notable sur la classification des locuteurs peu sévères, ce qui est observable avec l'entrée Mod_Cepstr du CNN apportant une *accuracy* de 18% pour les locuteurs les moins sévères en comparaison à la meilleure entrée du MLP (Meanch + SD Meanch).

Nous avons observé une importante différence d'*accuracy* minimum entre les entrées : pour la mesure Mod_Cepstr le locuteur le moins bien classé est à hauteur de 50% dans le groupe SCA, alors que, pour les entrées du MLP le minimum est entre 0 et 1% indiquant une très mauvaise classification pour au moins un des locuteurs SCA. Au sujet des CTRL, les entrées du CNN rencontrent davantage de difficultés par rapport à celles du MLP pour classifier correctement les locuteurs des groupes CTRL1 et CTRL2 sur l'ensemble des données soumises. Sa meilleure performance se trouve avec l'entrée Mod_Cepstr. Les entrées du MLP quant à elles classent mieux les CTRL que cela soit pour le test1 avec les CTRL1 ou le test2 (CTRL2). A contrario des entrées du CNN, deux entrées du MLP ont un minimum d'*accuracy* nettement plus élevé à 30%, 33% (toutes les entrées en entraînement et Mean pics) par rapport aux performances du minimum d'*accuracy* pour le CNN qui tournent autour de 0% pour l'intégralité de ses entrées. L'intégralité des entrées du MLP ont au moins un locuteur CTRL qui est classé à 100% pour le CNN ; nous avons ce taux uniquement pour deux entrées : Mod_Cepstr et Spectro montrant une fois de plus les difficultés du modèle sur cette population.

Au regard de l'écart-type de l'*accuracy* pour chaque entrée présent dans la Table 3, la mesure Mod_Cepstr a un écart-type parmi les plus bas, ce qui suggère une faible variation dans la classification des locuteurs SCA et une stabilité dans les performances. Pour la totalité des mesures données au MLP, nous observons de plus grands écarts-types indiquant une variabilité des classifications entre locuteurs, aussi bien pour le groupe SCA que les groupes CTRL1 et CTRL2.

4 Conclusion et Discussion

Nous avons observé des performances différentes des modèles CNN et MLP en fonction des types d'entrée. L'entrée Mod_Cepstr pour le groupe SCA s'est avérée meilleure face aux autres types d'entrées du CNN ciblées sur les informations de spectrogramme (Spectro, F1, 2, 3). Ces résultats s'observent pour les deux groupes CTRL. Ces performances peuvent s'expliquer par les multiples informations que porte la modulation cepstrale (Slis *et al.*, 2021; Lévêque *et al.*, 2022).

Le CNN avec des mesures non interprétables s'est montré plus performant pour la population SCA que le MLP avec des mesures issues d'une expertise phonétique (interprétables). Le MLP classe beaucoup mieux les patients très sévères que les patients peu sévères avec un écart-type qui montre une plus grande variation dans sa classification. L'efficacité de mesures non interprétables effectuées par un modèle de traitement automatique a également déjà été observée dans une étude antérieure où les mesures non interprétables ont surpassé celles étant interprétables (Favaro *et al.*, 2023). Nous avons observé que les mesures numériques sur la modulation cepstrale assimilées par le MLP sont plus performantes pour la population CTRL. Les difficultés du CNN sur la population CTRL pourraient s'expliquer par une variabilité moins prononcée ou différente au sein des locuteurs dans les groupes CTRL1 et CTRL2 par rapport aux SCA. Malgré les tests de plusieurs architectures de modèles et l'augmentation du volume de données pour cette catégorie durant l'entraînement, aucune amélioration des résultats n'a été observée. La quantité relativement faible de données disponibles dans cette étude pourrait également contribuer à ce problème. Il a été noté dans l'étude précédemment citée (Favaro *et al.*, 2023) que pour les corpus détenant plus de données, des mesures non interprétables (les traits provenant du modèle TRILLsson) étaient plus performantes et avaient un plus gros écart de performances en comparaison des mesures interprétables (des traits prosodiques).

Les difficultés du MLP à classer les locuteurs SCA avec une faible dysarthrie peuvent être dues à des mesures similaires entre CTRL et SCA dues à la précocité de la maladie (Slis *et al.*, 2021), ce qui peut causer des difficultés pour le MLP à apprendre un motif particulier pour cette population et à les discriminer convenablement. À contrario, l'aisance du MLP dans la classification de la population CTRL peut résider dans le fait d'avoir des mesures plus stables dans la population CTRL avec très peu de variations, ce qui peut l'aider à apprendre un pattern. La performance observée de l'entrée Meanch et SD Meanch démontre son importance (Slis *et al.*, 2021). Ces mesures peuvent refléter des troubles articulatoires comme des difficultés à atteindre des cibles articulatoires lors de transition d'un phonème à l'autre, ce qui est observable chez la population SCA (Shalling *et al.*, 2007). En effet, le modèle arrive à apprendre un pattern pour chaque classe, ce qui permet cette bonne classification. Globalement, avec la totalité des mesures condensées, il est observé un score inférieur de 8% du MLP par rapport au CNN avec l'entrée Mod_Cepstr pour la population SCA. L'entrée Mod_Cepstr dans le CNN peut capturer des motifs spatiaux et des relations locales dans les données qui ne sont pas évidentes ou directement accessibles via les mesures brutes de la modulation cepstrale utilisées dans le MLP. Cela peut inclure des nuances subtiles et des motifs complexes qui sont importants pour la classification.

Ces résultats laissent plusieurs ouvertures pour la suite : les mesures prises par le CNN sur la modulation cepstrale ont un avantage pour la population SCA, ce qui montre une possibilité d'affiner les mesures numériques ciblées par la suite pour pouvoir capter des caractéristiques clefs permettant de différencier un locuteur peu atteint d'un témoin. Des tests prometteurs pour la même tâche ont été réalisés à l'aide de grands modèles de langues pré-entrainés (w2v2) et cette piste est en cours d'exploration.

Références

ABDERRAZEK S., FREDOUILLE C., GHIO A., LALAIN M., MEUNIER C. & WOISARD V. (2020). Towards interpreting deep learning models to understand loss of speech intelligibility in speech disorders - step 1 : Cnn model-based phone classification. *Interspeech 2020*, **27**, 522–2526,. DOI : hal-03017394.

ALSHAMMRI R., ALHARBI G., ALHARBI E. & ALMUBARK I. (2023). Machine learning approaches to identify parkinson's disease using voice signal features. *Sec. Medicine and Public Health*, **6**. DOI : 10.3389/frai.2023.1084001.

AUDIBERT N. & FOUGERON C. (2012). Distorsions de l'espace vocalique : quelles mesures ? application à la dysarthrie. *JEP TALN*, **27**, 217–224. DOI : hal-02436294.

BERRAR D. (2019). Cross-validation. *Encyclopedia of Bioinformatics and Computational Biology*, p. 542–545. DOI : 10.1016/B978-0-12-809633-8.20349-X.

BOURBON A., FOUGERON C. & CREVIER_BUCHMAN L. (2023). *Effects of instruction and content on repetition performances of ataxic dysarthric and healthy speakers*. Thèse de doctorat, 8th International Conference on Speech Motor Control Groeningen.

BRENDEL B., SYNOFZIK M., ACKERMANN H., LINDIG T., SCHÖLDERLE T., SCHÖLS L. & ZIEGLER W. (2015). Comparing speech characteristics in spinocerebellar ataxias type 3 and type 6 with friedreich ataxia. *J Neurol*, **262**, 21–26. DOI : 10.1007/s00415-014-7511-8.

DARLEY F., ARONSON A. & BROWN J. (1969). Differential diagnostic patterns of dysarthria. *Journal of Speech and Hearing Research*, p. 246–269.

FARAGÓ P., ȘTEFĂNIGĂ S.-A., CORDOȘ C.-G., MIHĂILĂ L.-I. & HINTEA S. (2022). Cnn-based identification of parkinson's disease from continuous speech in noisy environments. *JSLHR*, **5**, 1767–1783. DOI : 10.3390/bioengineering10050531.

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 : 10.1016/j.compbiomed.2023.107559.

GOLDSTEIN L. (2019). The role of temporal modulation in sensorimotor interaction. *Front. Psychol*, **10**, 2068.

KIM L. & GENDROT C. (2022). Classification automatique de voyelles nasales pour une caractérisation de la qualité de voix des locuteurs par des réseaux de neurones convolutifs. *JEP*, p. 13–17.

KODRASI I., PERNON M., LAGANARO M. & BOURLARD H. (2021). Automatic and perceptual discrimination between dysarthria, apraxia of speech, and neurotypical speech. *ICASSP*.

LAARIDH I., FREDOUILLE C. & MEUNIER C. (2016). Détection automatique d'anomalies sur deux styles de parole dysarthrique : parole lue vs spontanée. *JEP*.

LAGANARO M., FOUGERON C., PERNON M., LEVÊQUE N., BOREL S., CHIUVÉ M. F. S. C., URSULA LOPEZ R. T., MÉNARD L., BURKHARD P. R., ASSAL F. & DELVAUX V. (2020). Sensitivity and specificity of an acoustic- and perceptual-based tool for assessing motor speech disorders in french : the monpage-screening protocol. *Clinical Linguistics & Phonetics*, **35**, 1060–1075. DOI : 10.1080/02699206.2020.1865460.

LÉVÊQUE N., SLIS A., LANCIA L., BRUNETEAU G. & FOUGERON C. (2022). Acoustic change over time in spastic and/or flaccid dysarthria in motor neuron diseases. *JSLHR*, **5**, 1767–1783. DOI : 10.1044/2022_JSLHR-21-00434.

MATHAD V. C., LISS J. M., CHAPMAN K., SCHERER N. & BERISHA V. (2022). Consonant-vowel transition models based on deep learning for objective evaluation of articulation. *IEEE*, **31**, 86–95. DOI : 10.1109/TASLP.2022.3209937.

SCHMITZ-HÜBSCH T., ECKERT O., SCHLEGEL U., KLOCKGETHER T. & SKODDA S. (2011). Instability of syllable repetition in patients with spinocerebellar ataxia and parkinson's disease. *Mov disord*, **27**, 316–319. DOI : 10.1002/mds.24030.

SHALLING E., HAMMARBERG B. & HARTELIUS L. (2007). Perceptual and acoustic analysis of speech in individuals with spinocerebellar ataxia (sca). *Logopedics Phoniatrics Vocology*, **32**, 31–46. DOI : 10.1080/14015430600789203.

SLIS A., FOUGERON C., LÉVÊQUE N., PERNON M., ASSAL F. & LANCIA L. (2021). Analysing spectral changes over time to identify articulatory impairments in dysarthria. DOI : 10.1121/10.0003332.

WANG J., KOTHALKAR P. V., CAO B. & HEITZMAN D. (2016). Towards automatic detection of amyotrophic lateral sclerosis from speech acoustic and articulatory samples. *Interspeech2016*. DOI : 10.21437/Interspeech.2016-1542.

XU L., LISS J. & BERISHA V. (2022). Dysarthria detection based on a deep learning model with a clinically-interpretable layer. *JASA*, **3**, 015201. DOI : 10.1121/10.0016833.