Interprétation des représentations profondes des traits phonétiques via l'approche NCD -Neuro-based Concept Detector : Application aux troubles de la parole

Sondes Abderrazek, Corinne Fredouille, Alain Ghio, Muriel Lalain, Christine Meunier, Virginie Woisard

Interprétation des représentations profondes des traits phonétiques via l'approche *NCD - Neuro-based Concept Detector* : Application aux troubles de la parole

Sondes Abderrazek[1] Corinne Fredouille[1] Alain Ghio[2]
Muriel Lalain[2] Christine Meunier[2] Virginie Woisard[3]
[1]LIA, Avignon University, France
[2]Aix-Marseille Univ, LPL, CNRS, Aix-en-Provence, France
[3]UT2J, Laboratoire de NeuroPsychoLinguistique, Toulouse University & Hospital, France
(sondes.abderrazek,corinne.fredouille)@univ-avignon.fr,
(alain.ghio,muriel.lalain,christine.meunier)@univ-amu.fr,
woisard.v@chu-toulouse.fr

RÉSUMÉ

La popularité des réseaux de neurones profonds (DNN) ne cesse de croître, tout comme l'intérêt de mieux comprendre leur fonctionnement. Dans ce papier, nous présentons un cadre de travail général, nommé *Neuro-based Concept Detector (NCD)*, pour interpréter les représentations profondes d'un DNN. Basé sur les schémas d'activation des unités cachées, ce cadre met en évidence la capacité des neurones à détecter un concept implicite en lien avec la tâche finale visée. Appliqué à la parole normale, le NCD révèle l'émergence de traits phonétiques (concept cible ici) dans les couches d'un modèle entraîné sur la tâche de classification des phonèmes français. Nous montrons également que, sur un corpus de cancers de la tête et du cou, les connaissances issues du NCD permettent de caractériser les troubles de parole des patients en terme d'altération des traits phonétiques, fournissant ainsi de premières informations très pertinentes pour la pratique clinique telle que la rééducation.

ABSTRACT

Interpreting deep representations of phonetic traits via the NCD approach - Neuro-based Concept Detector : Application to speech disorders

The popularity of deep neural networks (DNNs) continues to grow, as does the interest in better understanding how they work. In this paper, we present a general framework, called *Neuro-based Concept Detector (NCD)*, to interpret deep representations of a DNN. Based on the activation patterns of hidden units, this framework highlights the ability of neurons to detect an implicit concept related to the targeted final task. Applied to normal speech, the NCD framework reveals the emergence of phonetic features (target concept here) in the layers of a model trained on the French phoneme classification task. We further show that, on a corpus of head and neck cancers, the knowledge gained from NCD allows us to characterize the speech disorders of patients in terms of phonetic features alteration, thus providing first information very relevant for clinical practice such as rehabilitation.

MOTS-CLÉS : Apprentissage profond, Interprétabilité, Troubles de la parole, Traits phonétiques, Intelligibilité de la parole, Cancer de la tête et du cou.

KEYWORDS: Deep learning, Interpretability, speech disorders, phonetic traits, speech intelligibility, Heand and Neck Cancers (HNC).

1 Introduction

Les cancers de la tête et du cou (HNC) représentent le septième cancer le plus répandu dans le monde en 2018 avec 890 000 nouveaux cas diagnostiqués (Chow, 2020). Les fonctions respiratoires, de déglutition, de phonation et de parole peuvent être affectées par les HNC eux-mêmes, mais aussi par les soins thérapeutiques associés comme la chirurgie, la chimiothérapie et/ou la radiothérapie, qui peuvent être particulièrement mutilants. La communication des patients HNC avec autrui et les interactions socio-professionnelles qui en découlent peuvent être gravement affectées, pouvant réduire de manière drastique leur qualité de vie. Aussi, l'évaluation de la qualité de la parole est cruciale en pratique clinique compte tenu des impacts potentiels en cas de troubles (Woisard *et al.*, 2022). L'évaluation perceptive est l'outil le plus couramment utilisé dans la prise en charge des troubles de la parole, considérant les HNC, mais aussi les maladies neurodégénératives impliquant une dysarthrie (maladie de Parkinson, sclérose latérale amyotrophique, accidents vasculaires cérébraux, etc.), ainsi que la prise en charge des troubles de la voix en cas de dysphonie organique ou fonctionnelle. De nombreux protocoles, mesures ou échelles perceptives sont disponibles pour évaluer la qualité de la parole et de la voix dans les pratiques cliniques (Enderby, 1980; Auzou & Rolland-Monnoury, 2006; Darley *et al.*, 1975; Hirano, 1981). Néanmoins, si l'évaluation perceptive des troubles de la parole et de la voix est fréquente et bien documentée, une critique récurrente est la subjectivité des évaluations des cliniciens en raison de leur degré d'exposition à la parole pathologique mais également à leur habituation aux tests d'évaluation perceptifs, souvent jugés, par ailleurs, coûteux, longs et non-reproductibles. Pour faire face à ces limites, les approches automatiques sont apparues très tôt comme des solutions potentielles pour fournir des outils d'évaluation objectifs.

Concernant l'intelligibilité dans le cadre des troubles de la parole, objet de ce travail, un grand nombre d'études sont disponibles dans la littérature. Parmi elles, on peut distinguer différentes orientations de recherche : (1) les approches directement basées sur la transcription automatique de la parole et les taux d'erreurs mot comme score d'intelligibilité, comme réalisé par les cliniciens dans certains tests perceptifs, ou en extrayant des caractéristiques phonologiques issues des sorties ASR (Middag *et al.*, 2009; Tripathi *et al.*, 2020), (2) les approches basées sur l'extraction de caractéristiques spécifiques du signal de parole telles que des caractéristiques spectrales, articulatoires, prosodiques ou de qualité de la voix. Cette extraction est ensuite couplée à des approches classiques de classification ou de prédiction pour fournir un score d'intelligibilité telles que des modèles à base de mélanges de gaussiennes (GMM), les machines à vecteurs support (SVM), ou encore des modèles à base de réseaux profonds (Deep Neuronal Network - DNN), etc (Kim *et al.*, 2015; Orozco-Arroyave *et al.*, 2016; Narendra & Alku, 2021), (3) les approches spécifiques proposées pour l'évaluation de l'intelligibilité de la parole : à base d'i-vecteurs ou x-vecteurs (Martínez *et al.*, 2015; Quintas *et al.*, 2020), à base de détection d'anomalies (Laaridh *et al.*, 2015) ou de comparaison entre parole dégradée et signal de référence (Janbakhshi *et al.*, 2019), à base d'apprentissage profond comme les LSTM avec mécanisme d'attention (Fernández Díaz & Gallardo-Antolín, 2020).

Toutes ces études partagent le même objectif d'évaluer l'intelligibilité de la parole dans le contexte clinique en fournissant un score unique, tout en considérant une ou plusieurs dimensions de la parole. Si ces approches prennent en compte, souvent efficacement, les troubles de la parole et leur impact sur le signal de parole en matière d'altérations, elles sont finalement incapables de définir précisément le lien entre les troubles de la parole et le score d'intelligibilité qu'elles obtiennent. Pourtant, ce lien est important pour guider le clinicien dans son évaluation clinique.

L'objectif du travail présenté ici est d'apporter une solution intermédiaire dans le cadre d'un projet de recherche pluridisciplinaire dont l'objectif est la recherche des unités linguistiques jouant un rôle majeur dans l'intelligibilité de la parole, et notamment son altération lors de troubles de la parole. Dans

ce contexte, ce papier propose un cadre analytique général et original, nommé *Neuro-based Concept Detector* - NCD, spécialement conçu pour interpréter les représentations profondes d'un DNN. Basé sur les modèles d'activation des neurones cachés, ce cadre permet de mettre en évidence la capacité des neurones à détecter un concept cible en lien avec la tâche finale réalisée par le DNN, concept non représenté explicitement dans les données d'apprentissage. Appliqué à la phonétique clinique, ce cadre permet de fournir une modélisation efficace des caractéristiques acoustiques et articulatoires de la parole saine, au travers de détecteurs des traits phonétiques (concept cible ici), facilement interprétables en matière d'altérations dès lors que l'on considère des enregistrements présentant des troubles de la parole. Dans les étapes ultérieures (perspective de ce travail), nous tirerons parti des techniques d'apprentissage par transfert pour impliquer ce modèle et les connaissances associées (*détecteurs des traits phonétiques*) dans une tâche de prédiction d'intelligibilité. Ce dernier niveau, impliquant les connaissances extraites par l'approche NCD, devrait fournir une décision/explication plus éclairée concernant l'impact des altérations de la parole et leur nature phonétique sur l'évolution de l'intelligibilité de la parole des patients, et ainsi fournir *in fine* des informations pertinentes en plus d'une simple note d'évaluation de l'intelligibilité pour guider les cliniciens dans leur pratique.

2 Corpus utilisés et traits phonétiques du français

Comme mentionné en introduction, l'objectif de ce travail est de fournir une modélisation efficace des caractéristiques acoustiques et articulatoires de la parole saine, au travers de détecteurs de traits phonétiques (concept cible) en lien avec une architecture neuronale. Dans cet optique, le corpus BREF a été utilisé pour fournir l'ensemble des enregistrements de parole saine. Ce corpus est composé d'enregistrements de parole en français produits par 120 locuteurs mêlant hommes et femmes, sur une tâche de lecture de textes de journaux, représentant environ 115h de parole (Lamel *et al.*, 1991). Toutes les productions de parole ont été alignées automatiquement en utilisant un système d'alignement forcé, classiquement basé sur un algorithme de Viterbi et un modèle de Markov caché à trois états indépendant du contexte (modèle HMM) par phonème du français, entraîné sur des données distinctes de parole. Cet alignement fournit par conséquent les frontières temporelles de début et de fin de tous les phonèmes présents dans les signaux de parole.
Le corpus C2SI-LEC est une sous-partie du corpus de parole enregistrée dans le cadre du projet C2SI entre 2015 et 2017 (Woisard & et al., 2021). Le corpus global comprend des patients atteints de cancers de la tête et du cou (cavité orale ou oropharynx) et des contrôles. Tous les patients ont subi un traitement dédié consistant en une chirurgie et/ou radiothérapie et/ou chimiothérapie. Tous les locuteurs ont été enregistrés suivant différentes tâches de production de parole (tenue de voyelles /a/, pseudo-mots isolés, lectures de texte ou de phrases, description d'images et brève interview). Différentes évaluations perceptives ont été menées par un jury de 5 à 6 experts (phoniatres ou orthophonistes). Nous nous intéressons ici aux mesures de la sévérité de la parole (Sev-DESC) et de l'intelligibilité (Intel-DESC), sur une échelle de 0 (troubles majeurs) à 10 (aucun trouble) et à une mesure d'altération phonémique (Phon-DESC) sur une échelle de 0 (aucun trouble) à 3 (troubles majeurs), ces trois mesures perceptives étant évaluées à partir de la tâche de description d'images. Les notes données par les experts sont moyennées pour fournir une valeur unique par locuteur et par mesure perceptive. Dans cette étude, l'accent est mis sur la tâche de lecture uniquement, en considérant 89 enregistrements de parole produite par 82 patients (7 patients enregistrés deux fois) et 25 enregistrements pour 24 locuteurs contrôle (un locuteur contrôle enregistré deux fois). A partir du texte lu (systématiquement corrigé en cas d'erreurs de lecture), tous les signaux de parole ont été alignés automatiquement de manière similaire au corpus BREF.

	a	Ê	Û	Ô	u	y	i	ã	µ	ɔ̃
nasal	0	0	0	0	0	0	0	1	1	1
arrière		0	0	1	1	0	0	1	0	1
arrondi	0	0	1	1	1	1	0	0		1
haut	0	0	0	0	1	1	1	0	0	0
bas	1	0	0	0	0	0	0	1	0	0

	p	t	k	b	d	g	f	s	ʃ	v	z	ʒ	m	n	ɲ	l	R	j	w	ɥ
vocalique	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1		1	1	1
continu	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0		1	1	1	1
nasal	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
voisé	0	0	0	1	1	1	0	0	0	1	1	1	1	1	1	1		1	1	1
compact	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0	1	0	0
aigu	0	1	1	0	1	1	0	1	1	0	1	1	0	1	1	1	1	1	0	0

TABLE 1 – Déclinaison en traits phonétiques des voyelles et consonnes du français

Traits phonétiques des phonèmes du français

En phonologie, les phonèmes peuvent être décrits en caractéristiques acoustico-phonétiques plus élémentaires qui lient les représentations catégorielles abstraites des phonèmes aux propriétés acoustiques sous-jacentes et aux gestes articulatoires qui les génèrent. Différentes catégorisations existent pour définir l'ensemble des traits phonétiques composant et distinguant ces phonèmes. Dans cet article, nous avons adopté l'approche décrite dans (Ghio *et al.*, 2020) pour caractériser les phonèmes français en traits phonétiques en s'appuyant d'abord sur une séparation entre voyelles et consonnes (cf. table 1). Ici, la notion de traits phonétiques impose un statut binaire (égal à 1 si le trait phonétique est présent dans le phonème, 0 s'il est absent, indéfini si pas applicable). Chacune de ces caractéristiques acoustico-phonétiques décrit un élément particulier dans la manière dont le son est produit. Considérant le tableau 1, le trait phonétique *continu* est lié au mode d'articulation et oppose les occlusives *[-continu]* aux fricatives *[+continu]*. En effet, les occlusives sont produites par un blocage complet du flux d'air dans le conduit vocal suivi d'un phénomène bien caractéristique de libération de l'air comprimé, contrairement aux fricatives associées à un flux d'air s'échappant par un passage étroit dans la cavité buccale, émetteur du sifflement qui les caractérise. Le lieu d'articulation est quant à lui distingué, par exemple, par le trait phonétique *compact* pour lequel *[+compact]* cible des consonnes articulées contre le palais dur ou mou (vélaires et palatales) tandis que *[-compact]* cible les consonnes labiales et alvéolaires. En ce qui concerne les voyelles, il faut noter l'utilisation d'archiphonèmes [1] dans la représentation des traits phonétiques donnée dans le tableau 2, notamment Û={œ,ø}, Ô={o,ɔ}, Ê={e,ɛ} et µ={œ̃,ɛ̃}.

3 Cadre proposé : Neuro-based Concept Detector

Dans cette section, nous détaillons le cadre analytique général, NCD, proposé pour l'interprétation des unités cachées d'un DNN effectuant une tâche de classification. Dans une première partie, nous commençons par une brève description du modèle CNN auquel on applique l'approche NCD. Dans une deuxième partie, nous décrivons la méthodologie proposée pour qualifier chaque neurone et définir le sous-ensemble de neurones détecteurs d'un concept cible, à savoir, ici, le sous-ensemble de neurones interprétables détectant des caractéristiques phonétiques distinctives.

3.1 Tâche de classification des phonèmes du français

L'architecture à base de réseaux de neurones convolutifs (Convolutional Neural Network - CNN) utilisée dans cette étude a été détaillée dans nos précédents travaux (Abderrazek *et al.*, 2020).

1. unité phonologique qui exprime les traits communs de deux ou plusieurs phonèmes, impliqués dans une neutralisation.

L'apprentissage du modèle CNN pour une tâche de classification des phonèmes français repose sur le corpus BREF décrit dans la section 2. Un ensemble de 3M d'échantillons équilibré en phonèmes a été réparti en 90% pour l'apprentissage et 10% pour la validation du modèle. Ce modèle a atteint un taux de 80.8% de bonne classification sur des données de test BREF.

3.2 Définition des neurones détecteurs d'un concept cible : traits phonétiques

Représentation de l'activité des neurones

L'objectif ici est de caractériser l'activité des neurones lorsqu'ils sont exposés à la parole saine. Pour cela, un sous-ensemble représentatif de données BREF (différent de celui utilisé dans l'apprentissage du CNN) a été conçu. Ce sous-ensemble est équilibré en nombre de trames pour chacun des 30 phonèmes. Ainsi, pour la sélection de trames nous avons considéré l'ensemble des trames associées à un segment de parole (définies par l'alignement automatique) lié à une production complète d'un phonème (et ce pour tous les phonèmes considérés). Nous avons impliqué un maximum de contextes phonémiques et tous les locuteurs disponibles dans le corpus pour varier les productions de parole. Cette sélection a conduit à un sous-ensemble de données comprenant près de 82K de trames, que nous nommons *BREF-Int*.

La valeur d'activation du neurone n, compte tenu de la i^{ieme} trame d'entrée de notre jeu de données *BREF-Int* se définit par $h_{n,i}$. Une activation normalisée $a_{n,i}$ est calculée pour chaque neurone en divisant les valeurs d'activation initiales du neurone pour différentes trames d'entrée du jeu de données par le maximum atteint sur toutes ces valeurs : $a_{n,i} = \frac{h_{n,i}}{h_{max,n}}$ où $h_{max,n} = \max h_{n,j} \forall j$.

Un exemple de visualisation pour le neurone 214 de la 2e couche de classification (FC2) est fourni en figure 1 (figure de gauche) sous la forme d'un nuage de point (jitter plot). Ce neurone a été identifié grâce à l'approche automatique décrite ci-après pour désigner les neurones détecteurs de traits phonétiques cibles. Chaque point du graphique correspond à l'activation normalisée de cette unité en réponse à une trame de *BREF-Int*, cette trame étant associée à un phonème donné, nommé en ordonnée. Comme on peut l'observer, ce neurone a une réponse distincte pour les trois consonnes nasales (/n/,/m/,/ɲ/) - ensemble de points encerclé.

Caractérisation des neurones détecteurs de concept

En vue d'aligner chaque neurone avec les traits phonétiques définis en section 2 et de mesurer le degré d'encodage des unités cachées pour chacun de ces traits, un score doit être défini. Ce score doit quantifier, pour chaque neurone, le degré de détection de la présence d'un trait phonétique en fonction de l'activation des phonèmes associés à ce trait phonétique, et de manière complémentaire, des phonèmes qui ne le présentent pas. Ce score va reposer sur les valeurs d'activations normalisées pour les unités individuelles, présentées dans la section précédente pour chacun des 30 phonèmes.

Pour chaque neurone n, $A_{n,k}^{BREF}$ est l'ensemble des activations normalisées du neurone *n* pour toutes les trames ayant pour label le phonème *k* et appartenant à *BREF-Int*. Nous notons la valeur médiane d'activation du neurone n pour le phonème k comme $m_{A_{n,k}^{BREF}}$. Le score S_{n,T_x}, quantifiant le degré avec lequel une unité détecte la présence/absence d'un trait phonétique, est donc calculé pour chaque neurone n et trait phonétique T_x comme suit :

$$S_{n,T_x} = \frac{1}{|[+T_x]|} \sum_{k \in [+T_x]} m_{A_{n,k}^{BREF}} - \frac{1}{|[-T_x]|} \sum_{k \in [-T_x]} m_{A_{n,k}^{BREF}} \quad (1)$$

Les traits phonétiques étant des concepts binaires caractérisant séparément les voyelles et les consonnes, cette distinction est intégrée au score grâce à $x \in [v, c]$, qui désigne la macro-classe des voyelles ou des consonnes, respectivement v et c. Par conséquent, $T_v \in$

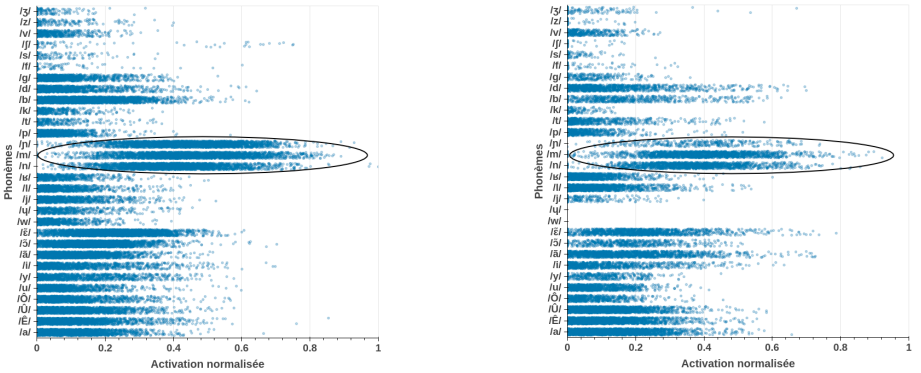

FIGURE 1 – Nuages de points illustrant les activations normalisées pour l'unité 214 de FC2. A gauche, corpus *BREF-Int*, à droite locuteurs contrôle de *C2SI-LEC*.

{nasal, arrière, haut, arrondi, bas} et $T_c \in$ *{vocalique, continu, nasal, voisé, compact, aigu}*.
Il est important de mentionner que le score $S_{n,T_x} \in [-1; 1]$. Ainsi, une valeur proche de 1 reflète que le neurone est un détecteur fort de la présence du trait phonétique en question puisqu'il distingue les phonèmes présentant les caractéristiques par un niveau d'activation élevé. Dans le même temps, un niveau d'activation très faible représente l'ensemble complémentaire des phonèmes ne présentant pas le trait. À l'inverse, lorsqu'un neurone a un score très faible proche de -1, cela signifie que le neurone est un détecteur fort du trait phonétique opposé, ce qui est également pertinent dans une potentielle analyse. Sur cette base, nous considérons que le neurone n est un détecteur de la présence du trait phonétique T_x, noté [+T_x], si S_{n,T_x} dépasse un seuil. A l'inverse, si S_{n,T_x} est inférieur au seuil, le neurone n est considéré comme un détecteur de l'absence du trait phonétique T_x, noté [-T_x]. De toute évidence, différents seuils pourraient conduire à un nombre différent de neurones sélectionnés comme détecteurs de traits phonétiques à travers les couches. Cependant, nous observons qu'il n'en résulte pas de changement significatif en terme de répartition de cet ensemble de neurones sur les différents traits phonétiques. Ainsi, nous avons empiriquement fixé le seuil à $\pm 0,25$ [2]. Etant donné qu'un neurone peut être détecteur de plusieurs traits phonétiques (associés à des scores pertinents respectant le seuil), le trait phonétique présentant le meilleur score est choisi dans ce cas. Si le neurone est identifié comme détecteur à la fois pour les macro-classes voyelles et consonnes, il sera considéré comme détecteur du trait phonétique présentant le meilleur score pour chacune d'elles.

4 Application de l'approche NCD aux troubles de la parole

Définition du score de similarité

Dans cette section, il s'agit de voir comment l'approche NCD décrite précédemment peut être utilisée dans le contexte des troubles de la parole et apporter de l'information utile en terme de traits phonétiques impactés. Observons à nouveau la figure 1 (figure de droite cette fois-ci) qui représente les activations normalisées, toujours pour le neurone 214 de la 2e couche de classification, associées aux trames des productions de parole de l'ensemble des contrôles du corpus *C2SI-LEC*. Malgré les différences entre les données - le corpus C2SI-LEC ne partage ni les mêmes conditions (matériel d'enregistrement et emplacement) ni les mêmes locuteurs que le corpus BREF sur lequel le modèle CNN a été entraîné - le neurone 214 a conservé strictement le même comportement global et par

2. L'intervalle de scores $[-1; 1]$ est théorique. Les expériences empiriques montrent plutôt un intervalle de valeurs entre $[-0.5; 0.5]$, d'où cette valeur seuil fixe.

TABLE 2 – Scores de corrélation entre les scores de similarité moyennés par trait sur l'ensemble des locuteurs du corpus *C2SI-LEC* pour les couches FC2 et FC3 et les mesures perceptives associées.

		Sev-DESC	Intel-DESC	Phon-DESC
FC2 / FC3	**Voyelles**	0.84 / 0.83	0.74 / 0.72	-0.80 / -0.77
	Consonnes	0.90 / 0.89	0.82 / 0.81	-0.86 / -0.85

conséquent sa capacité à détecter le trait phonétique *[+nasal]* sur les consonnes.
Dans cette optique, considérons N_t l'ensemble des neurones interprétables sélectionnés comme détecteurs pour le trait phonétique *t* sur le corpus *BREF-Int*. Considérons maintenant une séquence de parole d'un locuteur issu du corpus *C2SI-LEC*, l'application de l'approche NCD consiste à observer la réponse évoquée de tous les neurones appartenant à N_t pour les seuls phonèmes de la séquence de parole présentant le trait phonétique *t*. Ainsi, à l'instar de $A_{n,k}^{BREF}$ introduit dans la section 3.2, on note $A_{n,k}^{s}$ l'ensemble des activations normalisées du neurone *n* pour toutes les trames appartenant au phonème *k* et produites par le locuteur *s* du corpus C2SI-LEC. De même, on note $m_{A_{n,k}^{s}}$ la valeur médiane de cet ensemble d'activations normalisées, permettant de définir le rapport de score suivant :

$$\Theta_{s,t} = \frac{\sum_{n \in N_t} \sum_{k \in t} m_{A_{n,k}^{s}}}{\sum_{n \in N_t} \sum_{k \in t} m_{A_{n,k}^{BREF}}} \tag{2}$$

Il convient de noter que ce score peut être décliné en fonction des couches observées avec $N_t = N_{L,t}$ ou estimé globalement considérant l'ensemble des neurones détecteurs toutes couches confondues.

Score global par locuteur

Le score $\Theta_{s,t}$ défini ci-dessus permet d'évaluer le niveau de production moyen d'un trait phonétique donné par un locuteur, basé sur l'ensemble des neurones détectant le trait phonétique en question. Ce score varie de zéro à une valeur maximale indéfinie. Une valeur supérieure à 1 n'apporte pas plus d'information qu'une production parfaite du trait phonétique par le locuteur. Elle a donc été plafonnée à 1. Au contraire, un score proche de 0 signifie que quasiment aucun des détecteurs de traits phonétiques n'a expressément fourni une réponse sélective pour les phonèmes présentant ce trait, produits par le locuteur. Dans ce cas, nous pouvons supposer que la production de parole du locuteur ne présente pas de caractéristiques acoustiques typiques, mais plutôt une parole sévèrement altérée. Pour confirmer cette hypothèse, des analyses basées sur les coefficients de corrélation de Pearson ont été menées entre les scores de similarité $\Theta_{s,t}$, moyennés sur l'ensemble des traits acoustiques pour les couches de classification FC2 et FC3 pour chacun des locuteurs du corpus *C2SI-LEC* et les mesures perceptives associées. Fournies dans la table 2, nous pouvons observer de très bonnes corrélations, avec notamment un excellent score de 0.9 pour la mesure de sévérité considérant la macro-classe des consonnes. Ces corrélations montrent que le score de similarité global par locuteur, moyenné sur l'ensemble des traits acoustiques, permet de prendre en compte les différents niveaux de dégradation de la parole chez les patients du corpus *C2SI-LEC*.

Scores par trait phonétique et par locuteur du corpus *C2SI-LEC*

L'attention est portée ici sur chaque trait phonétique considéré individuellement. Au vu de nos objectifs initiaux en lien avec l'intelligibilité de la parole, il est d'un grand intérêt d'étudier les traits phonétiques des patients, leur possible altération, et de mettre ces connaissances en relation avec les troubles de parole de ces derniers. Dans ce contexte, on s'intéresse ici aux scores locaux évaluant individuellement dans quelle mesure un trait phonétique distinct de la macro-classe des consonnes a été produit par un locuteur spécifique. Ces scores locaux ont été projetés sous forme de carte thermique dans la figure 2 (la même analyse a été réalisée pour les voyelles, non présente ici faute de place). L'axe des abscisses représente les locuteurs du corpus *C2SI-LEC* triés par mesure perceptive de sévérité (du plus sévère à gauche au moins sévère à droite), tandis que l'axe des ordonnées représente

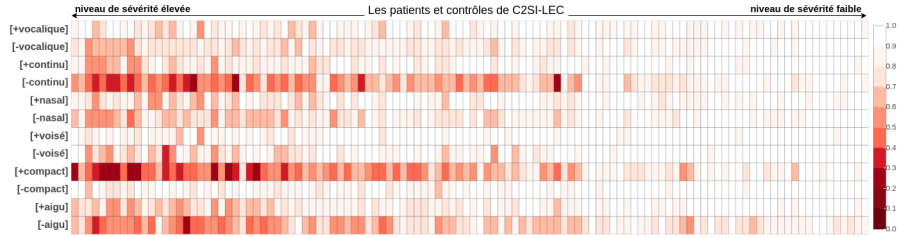

FIGURE 2 – Variation des scores par trait phonétique des consonnes (ordonnée) et par locuteur *C2SI-LEC* (abscisse) triés par mesure de sévérité *Sev-DESC* - à gauche niveau le plus sévère

les traits phonétiques relatifs aux consonnes. Une échelle séquentielle montre la variation des valeurs des scores faibles à élevées. Comme première observation, nous pouvons clairement mentionner que les cellules à forte opacité sont concentrées à gauche de la carte, comportement cohérent puisque cette partie du graphe correspond aux patients dont les troubles sont les plus sévères.
En considérant les traits phonétiques individuellement, nous pouvons observer en premier lieu que le trait de voisement n'est quasiment pas affecté, même pour les patients présentant une mesure de sévérité très élevée. Cette observation est cohérente, étant donné que les patients du corpus C2SI ne souffrent pas de cancer du larynx et n'ont donc pas a priori de raison de présenter une altération de ce trait. A l'inverse, il est intéressant d'observer les variations de scores du trait *[+compact]* suivant les locuteurs. En effet, ce trait se caractérise par une forte mobilisation de la langue, organe particulièrement touché pour un certain nombre de patients du corpus. La même observation peut être faite à partir de la carte thermique des voyelles (non présentée ici) sur laquelle les traits phonétiques *[+haut]* et *[+arrière]* sont les plus altérés chez les patients les plus atteints, traits marqués également par une forte mobilisation de la langue.

5 Conclusion

Ce papier présente un cadre général basé sur l'apprentissage profond et l'activité neuronale dont l'objectif est de mettre en évidence un sous-ensemble d'unités neuronales désignées comme détecteur d'un type spécifique de connaissances (Neuro-based concept Detector). Ici, l'accent est mis sur les traits phonétiques des phonèmes du français. Les expériences menées montrent que : (1) nous sommes capables de définir un ensemble de détecteurs de traits phonétiques sur la parole saine ; (2) sur la base de cette sélection de détecteurs de traits phonétiques, le score proposé, spécialement conçu pour rendre compte de l'écart entre la référence (ici la parole saine) et les nouvelles données (ici la parole altérée) est fortement corrélé aux mesures perceptives lorsqu'il est considéré globalement par locuteur, tout trait confondu ; (3) la prise en compte des scores de similarité locaux - un score par locuteur et par trait phonétique - mène à des observations très intéressantes, notamment le rôle de la langue pointé comme prédominant par l'analyse locale des traits phonétiques. L'ensemble de ces résultats sont très prometteurs pour la mise en œuvre des dernières étapes du projet, le transfert du modèle et des connaissances inhérentes pour la tâche de prédiction de l'intelligibilité et la mise en évidence des unités linguistiques impliquées dans sa variation en cas de troubles.

Remerciements

Ce travail est financé par l'ANR dans le cadre du projet RUGBI (nºANR-18-CE45-0008-04).

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. Shanghai, China.

AUZOU P. & ROLLAND-MONNOURY V. (2006). *Batterie d'évaluation clinique de la dysarthrie*. Édition Ortho.

CHOW L. Q. (2020). Head and neck cancer. *New England Journal of Medicine*. PMID : 31893516.

DARLEY F. L., ARONSON A. E. & BROWN J. R. (1975). *Motor speech disorders*. Philadelphia : W. B. Saunders and Co.

ENDERBY P. (1980). Frenchay dysarthria assessment. *British Journal of Disorders of Communication*.

FERNÁNDEZ DÍAZ M. & GALLARDO-ANTOLÍN A. (2020). An attention long short-term memory based system for automatic classification of speech intelligibility. *Engineering Applications of Artificial Intelligence*.

GHIO A., LALAIN M., GIUSTI L., FREDOUILLE C. & WOISARD V. (2020). How to compare automatically two phonological strings : Application to intelligibility measurement in the case of atypical speech. In *12th Conference on Language Resources and Evaluation (LREC)*, France.

HIRANO M. (1981). Psycho-acoustic evaluation of voice : GRBAS scale for evaluating the hoarse voice. *Clinical Examination of voice, Springer Verlag*.

JANBAKHSHI P., KODRASI I. & BOURLARD H. (2019). Pathological speech intelligibility assessment based on the short-time objective intelligibility measure. In *ICASSP'19*, UK.

KIM J., KUMAR N., TSIARTAS A., LI M. & NARAYANAN S. S. (2015). Automatic intelligibility classification of sentence-level pathological speech. *Computer Speech Language*.

LAARIDH I., FREDOUILLE C. & MEUNIER C. (2015). Automatic detection of phone-based anomalies in dysarthric speech. *ACM Transactions on Accessible Computing*, **6**(3).

LAMEL L. F., GAUVAIN J. L. & ESKÉNAZI M. (1991). BREF, a large vocabulary spoken corpus for French. In *Eurospeech'91*, Italy.

MARTÍNEZ D., LLEIDA E., GREEN P., CHRISTENSEN H., ORTEGA A. & MIGUEL A. (2015). Intelligibility assessment and speech recognizer word accuracy rate prediction for dysarthric speakers in a factor analysis subspace. *ACM Transactions on Accessible Computing (TACCESS)*, **6**(3), 10.

MIDDAG C., MARTENS J.-P., NUFFELEN G. V. & BODT M. D. (2009). Automated intelligibility assessment of pathological speech using phonological features. *EURASIP Journal on Applied Signal Processing*, **2009(1)**.

NARENDRA N. & ALKU P. (2021). Automatic assessment of intelligibility in speakers with dysarthria from coded telephone speech using glottal features. *Computer Speech Language*, **65**.

OROZCO-ARROYAVE J., VDSQUEZ-CORREA J., ARIAS-LONDO J., VARGAS-BONILLA J., SKODDA S., RUSZ J., NOTH E. *et al.* (2016). Towards an automatic monitoring of the neurological state of parkinson's patients from speech. In *ICASSP'16*, China.

QUINTAS S., MAUCLAIR J., WOISARD V. & PINQUIER J. (2020). Automatic Prediction of Speech Intelligibility Based on X-Vectors in the Context of Head and Neck Cancer. In *Interspeech*, China.

TRIPATHI A., BHOSALE S. & KOPPARAPU S. K. (2020). A novel approach for intelligibility assessment in dysarthric subjects. In *ICASSP'20*, Spain.

WOISARD V., BALAGUER M., FREDOUILLE C., FARINAS J., GHIO A., LALAIN M., PUECH M., ASTESANO C., PINQUIER J. & LEPAGE B. (2022). Construction of an automatic score for the evaluation of speech disorders among patients treated for a cancer of the oral cavity or the oropharynx : The carcinologic speech severity index. *Head & Neck*.

WOISARD V. & ET AL. (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)**.