Utilisation de modèles transformers pour la prédiction de l'intelligibilité de la parole de patients atteints de cancers des voies aérodigestives supérieures

Sebastião Quintas[1] Alberto Abad[2] Julie Mauclair[1]
Virginie Woisard[3] Julien Pinquier[1]
(1) IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
(2) INESC-ID, Instituto Superior Técnico, Lisbonne, Portugal
(3) CHU Larrey, Oncopole, Toulouse, France
{sebastiao.quintas, julie.mauclair, julien.pinquier}@irit.fr
alberto.abad@inesc-id.pt, woisard.v@chu-toulouse.fr

RÉSUMÉ

La prédiction automatique de l'intelligibilité de la parole est un sujet pertinent pour l'évaluation de la parole pathologique. Cette automatisation permet de limiter les biais et la subjectivité liés aux évaluations perceptuelles, évaluations très courantes en pratique clinique, et ainsi rendre plus robuste l'évaluation du patient. Dans ce travail, nous calculons un score d'intelligibilité basé sur une tâche de décodage acoustique-phonétique, où un ensemble de pseudo-mots est prononcé par des patients atteints de cancer des voies aérodigestives supérieures. Notre approche est fondée sur des réseaux de neurones profonds avec attention. Nous obtenons un score automatique d'intelligibilité fortement corrélé à l'évaluation perceptive ($\rho = 0,87$). En étudiant la fiabilité de notre prédiction sur un sous-ensemble de notre corpus, nous montrons qu'il est possible d'obtenir des résultats très similaires en utilisant de faibles quantités de données.

ABSTRACT

Using Transformers for the Automatic Prediction of Speech Intelligibility in the Context of Head and Neck Cancers

The automatic prediction of speech intelligibility is a widely relevant topic, specially when considering the clinical applications that it has within pathological speech. Due to the characteristic bias, variance and subjectivity associated with the perceptual evaluations, the standard in clinical contexts, an automatic approach becomes relevant. In the present work we aim to devise an automatic way to regress an intelligibility score based on acoustic-phonetic decoding from a set of pseudo-words from head and neck cancers. Our suggested approach is based on deep neural networks with an attention mechanism (transformers), and it presents not only a high correlation value of 0.87. Moreover, on the present work we study the reliability of using smaller amounts of data in the automatic prediction, showing that it is possible to obtain very similar results when using drastically smaller subsets of the original data.

MOTS-CLÉS : Parole pathologique, intelligibilité de la parole, réseau de neurones profond, transformer, cancer des voies aérodigestives supérieures.

KEYWORDS: Pathological Speech, Speech Intelligibility, Deep Neural Network, Transformer, Head and Neck Cancer.

10.21437/JEP.2022-7

1 Introduction

Les troubles de la parole, comme la dysarthrie ou la dysphonie, sont très souvent liés à des conditions médicales sous-jacentes. Ces troubles, conséquences de maladies telles que la sclérose latérale amyotrophique, Parkinson et les cancers des voies aérodigestives supérieures, peuvent affecter de multiples composantes de la parole (respiration, articulation, phonation, etc.) et peuvent provoquer différentes sortes des problèmes d'élocution. Les méthodes d'évaluation du trouble de la parole peuvent être généralistes ou liées à des maladies spécifiques. Les cancers des voies aérodigestives supérieures ont des répercussions fonctionnelles majeures sur la respiration, la déglutition et la phonation. À cause de cela, une altération de la communication est susceptible d'apparaître, impactant la qualité de vie du patient. Afin de régler ce problème, les évaluations perceptives sont très importantes dans le contexte clinique pour bien juger de l'évolution du patient et pour adapter son suivi. Néanmoins, ces évaluations sont chronophages et subjectives, notamment à cause de l'habituation de ce type de voix par le praticien, affectant la reproductibilité du score donné par celui-ci (Balaguer *et al.*, 2019). Pour cette raison, une approche automatique a été considérée comme une alternative plus rapide et plus objective par rapport aux méthodes perceptives (Middag, 2012).

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

Dans le cadre des prédictions automatiques de l'intelligibilité de la parole, nous pouvons distinguer des approches différentes. Ces approches peuvent aller de la régression d'un score d'intelligibilité à partir d'un taux d'erreur mot obtenu par un système de reconnaissance automatique de la parole (Christensen *et al.*, 2012), à l'extraction de paramètres pertinents d'une parole pathologique, en utilisant des technologies de traitement automatique de la parole (Quintas *et al.*, 2020). Étant donné que les approches basées sur le taux d'erreur mot sont moins performantes sur les patients sévères, et que les approches basées sur le traitement automatique de la parole sont normalement plus difficiles à interpréter, dans ce travail, nous proposons une méthode automatique pour prédire l'intelligibilité de la parole fondée sur le score individuel de plusieurs pseudo-mots énoncés par un locuteur. Le score final est ainsi calculé en fonction des scores individuels des différents mots prononcés par chaque patient.

Étant donné que la fatigue du patient est un problème récurrent dans l'enregistrement des tâches de parole (elle peut conduire à laisser les tâches incomplètes), nous voulons également évaluer comment les méthodes proposées se comporteront lors de l'utilisation de plus petites quantités de données, qui correspondent dans le cas présent à une plus petite quantité de pseudo-mots utilisés au moment de l'inférence.

Le reste de ce papier est organisé ainsi : la section 2 présente la méthodologie utilisée dans le cadre de ce travail, la préparation des données, la modélisation, puis la régression. La section 3 affiche nos résultats en utilisant le corpus français du cancer des voies aérodigestives supérieures (C2SI) (Woisard *et al.*, 2020). La section 4 analyse nos résultats, et propose une discussion.

2 Méthodologie

La méthodologie proposée pour faire la prédiction automatique de l'intelligibilité, repose sur l'utilisation d'un transformer avec un système d'attention. Nous pouvons diviser le système en 3 parties distinctes. La première correspond à la **préparation des données** et l'extraction des paramètres. La deuxième partie est la **modélisation** : nous utilisons un transformer afin d'obtenir les scores automatiques au niveau de chaque pseudo-mot. Finalement, la troisième partie est une **régression** du score général pour chaque locuteur en fonction des scores individuels de chaque mot. La figure 1 illustre la chaîne de traitement de notre système.

2.1 Préparation des données

La première partie du modèle correspond à l'extraction des paramètres. En entrée du système, nous avons les enregistrements des pseudo-mots des différents locuteurs, chaque mot a été enregistré individuellement. À partir de ces fichiers audios, nous calculons sur chaque fenêtre 40 banque de filtres (filterbanks). La taille et le pas de la fenêtre utilisée sont respectivement 25ms et 10ms. Chaque mot est associé à son score respectif de décodage acoustico-phonétique perceptuel. Ces scores sont utilisés comme les scores perceptifs de référence. La description de l'obtention de ces scores est décrite plus en détail dans la sous-section 3.1.

2.2 Modélisation

La deuxième partie du modèle correspond au système proposé pour ce travail : Le transformer avec mécanisme d'attention. Après l'obtention des filterbanks, les paramètres sont transmis à un transformer. Ce type de modèle, proposé par (Vaswani *et al.*, 2017) et adapté à la reconnaissance vocale par (Dong *et al.*, 2018), suit une architecture encodeur-décodeur. Notre proposition pour le transformer utilise un encodeur récurrent bidirectionnel avec des GRU (Gated Recurrent Units). L'encodeur possède 3 couches récurrentes avec une taille d'entrée de 40 (dimension des filterbanks) et une dimension cachée de 100. La sortie de l'encodeur est suivie d'un mécanisme d'attention. Ce mécanisme permet au système de se concentrer davantage sur des parties particulières du fichier d'entrée, tout en ignorant les parties moins pertinentes. Avec cela, nous espérons que le système apprendra automatiquement des interdépendances intéressantes entre des phonèmes consécutifs et qu'il trouvera le lien entre les erreurs de prononciation et le score DAP de chaque mot. Après le mécanisme d'attention, le vecteur de longueur fixe passe à travers un ensemble de 3 couches entièrement connectées, de dimension $[100 * 100]$ avec des ReLUs (Rectified Linear Units) comme fonctions d'activation. Enfin, nous utilisons une couche de Global Max Pooling pour obtenir le score individuel pour chaque mot et des connexions saute-couches entre la sortie du mécanisme d'attention et la couche de Global Max Pooling.

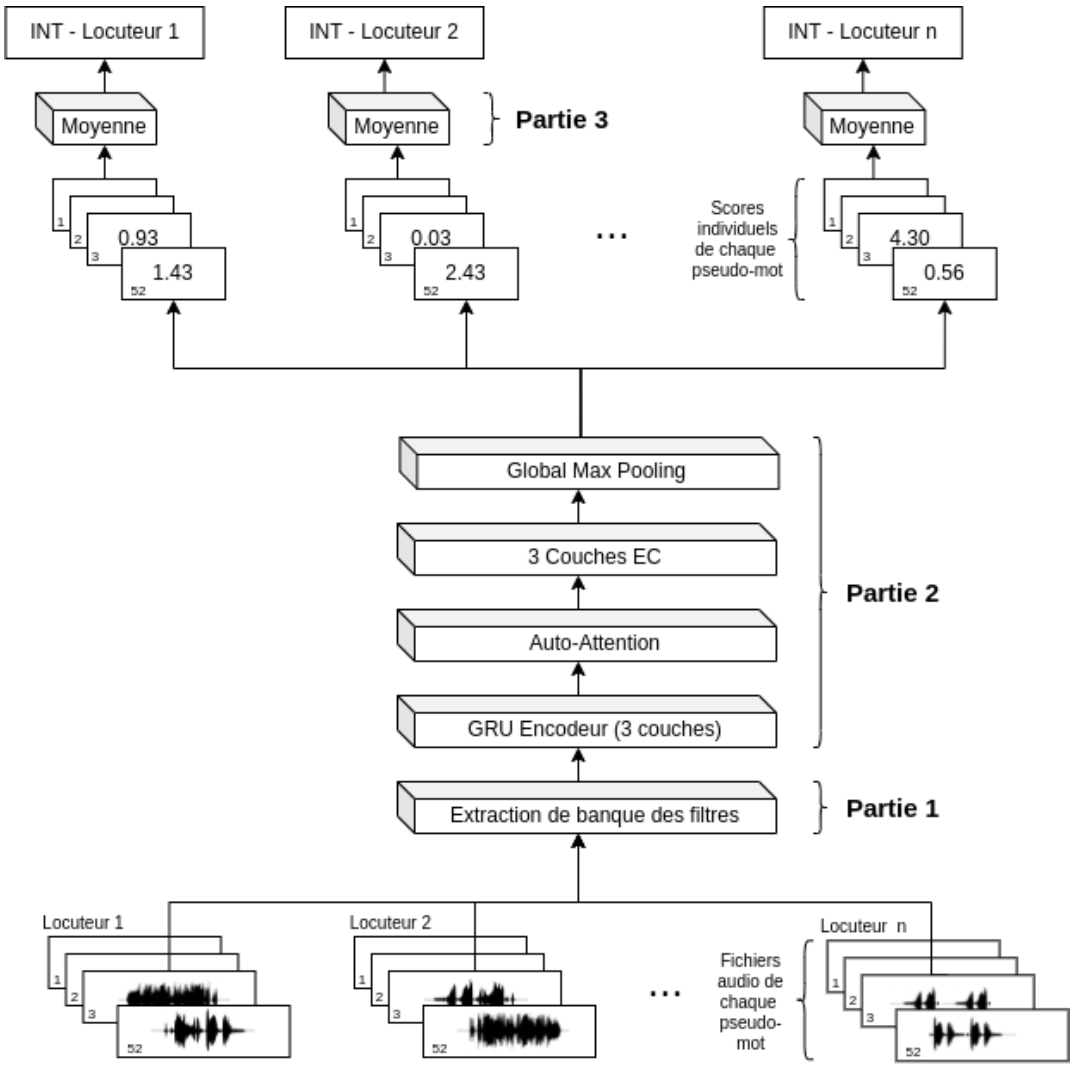

FIGURE 1 – Aperçu général du système proposé. EC signifie entièrement connectés. INT signifie le score d'intelligibilité d'un locuteur donné

2.3 Régression automatique du score du DAP

La troisième et dernière partie de notre modèle correspond à l'obtention du score d'intelligibilité pour chaque locuteur, en fonction du score de ses pseudo-mots. À partir des scores automatiques individuels d'intelligibilité de chaque pseudo-mot, nous calculons un score d'intelligibilité global pour chaque patient. Ici, nous avons utilisé la moyenne des scores individuels des 52 pseudo-mots de chaque locuteur.

En utilisant la moyenne, le système régresse un score basé sur le traitement automatique d'une variété de pseudo-mots avec différentes co-articulations entre les consonnes et les voyelles, rendant le score automatique riche et diversifié en termes de contenu phonétique.

3 Expériences et Résultats

3.1 Corpus C2SI

Dans le cadre de ce travail, le corpus français C2SI (Woisard *et al.*, 2020) a été utilisé. Celui-ci comprend des patients qui souffrent de cancer de la cavité buccale et de l'oropharynx ainsi que des locuteurs sains. Tous les locuteurs ont été invités à enregistrer un ensemble de tâches orales telles que une génération d'un /a/ tenu, une description d'image, un discours spontané, des lectures de texte et de pseudo-mots isolés.

Pour cet article, les études ont porté sur la tâche de lecture de pseudo-mots. Dans le contexte du corpus C2SI, tous les locuteurs ont enregistré 52 pseudo-mots différents qui respectent l'orthographe et la prononciation françaises (Ghio *et al.*, 2018). L'ensemble de 52 mots était différent pour chaque patient, suivant la structure suivante : $C(C)_1V_1C(C)_2V_2$, où $C(C)_i$ est une consonne isolée ou un groupe consonantique et V_i est une voyelle. Le tableau 1 présente quelques exemples des pseudo mots utilisés. Chaque ensemble de 52 mots comporte un sous-ensemble de 16 mots avec double consonne au début, 16 mots avec double consonne au milieu et au moins 26 mots sans double consonne. Les mots peuvent avoir des doubles consonnes à la fois au début et au milieu.

TABLE 1 – Exemples d'un ensemble de 52 pseudo-mots. Le bleu (resp. le violet) correspond aux doubles consonnes en début (resp. en milieu) de pseudo-mots.

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

La mesure perceptuelle d'intelligibilité utilisée a été obtenue en moyennant le score de la transcription individuelle de chaque pseudo-mot par 3 auditeurs naïfs. Ce score perceptif est utilisé ici comme vérité terrain (référence). Il a été calculé en fonction de la distance entre le mot transcrit et le mot d'origine, en fonction d'une matrice de coût des voyelles et des consonnes (Ghio *et al.*, 2018). Dans le cadre de ce travail, les scores perceptifs étaient compris entre 0, correspondant aux mots parfaitement intelligibles et 5, correspondant aux mots non-intelligibles. La moyenne des 3 auditeurs correspond au score de chaque mot, et la moyenne des 52 mots de chaque patient correspond au score DAP de chaque patient respectivement. Le même ensemble de 126 locuteurs (40 contrôles et 86 patients) que dans (Fredouille *et al.*, 2019) est utilisé.

3.2 Entraînement du Système

Pour entraîner notre système, nous avons utilisé une validation croisée à 10 blocs. À chaque bloc, 113 locuteurs (patients et contrôles) sont utilisés pour l'entraînement, et 13 locuteurs pour l'évaluation. Pour chaque bloc, nous avons entraîné le système pendant 200 itérations (epochs). Nous avons utilisé

un schéma d'ordonnancement du taux d'apprentissage avec décroissance polynomiale, à partir de 0,001 jusqu'à 0,0001 sur les 50 premières itérations. Des lots de taille 16 (batch size) ont été utilisés pendant l'entraînement et l'optimiseur Adam a été privilégié.

3.3 Évaluation du Système

La corrélation de Spearman (ρ) ainsi que l'erreur quadratique moyenne ($RMSE$, Root Mean Squared Error) nous ont permis d'évaluer notre système. La cible étant les scores perceptuels de DAP mentionnés précédemment en sous-section 3.1.

Nos résultats sont présentés dans le tableau 2 et illustrés sur la figure 2. Ils sont comparés avec une autre approche, provenant de la transcription automatique des pseudo-mots (Fredouille *et al.*, 2019; Ghio *et al.*, 2018). Celle-ci a obtenu les scores de DAP en utilisant un algorithme Wagner-Fischer entre les scores de référence et les scores obtenus de la transcription automatique. En utilisant les mêmes locuteurs et les mêmes pseudo-mots, nous obtenons de meilleurs résultats : sept points de plus de corrélation et la mesure d'erreur (RMSE) est réduite de près de la moitié.

TABLE 2 – Comparaison entre les résultats de référence et les résultats obtenus avec notre approche.

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

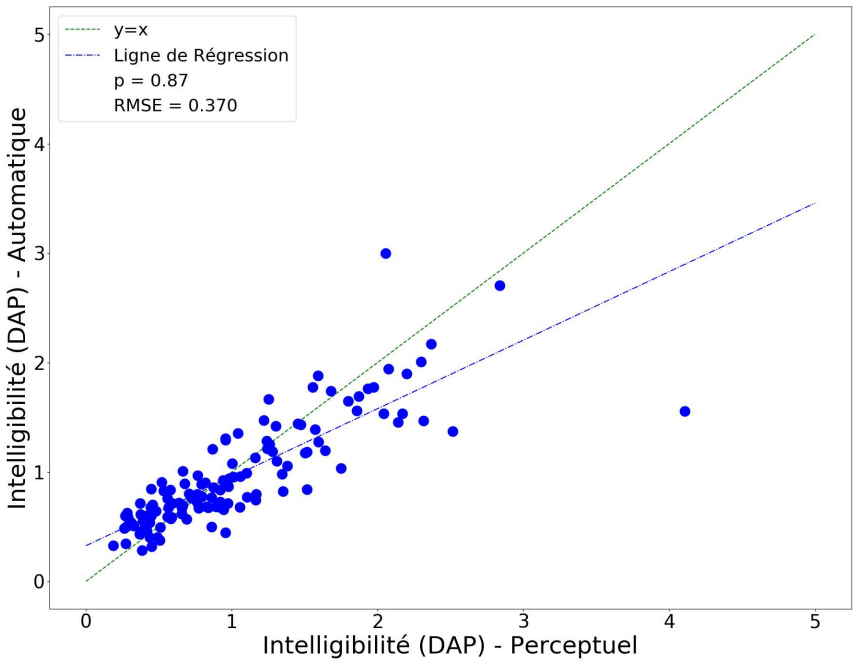

FIGURE 2 – Graphique des résultats obtenus, pour la prédiction automatique d'intelligibilité.

4 Analyse des résultats

4.1 Réduction de la quantité des pseudo-mots

Dans le cadre de cette étude, nous avions à notre disposition des ensembles de 52 pseudo-mots pour obtenir une mesure automatique d'intelligibilité. Cependant, dans un contexte clinique, l'enregistrement de 52 mots est chronophage. À cause de cela, il devient pertinent d'analyser s'il est possible d'obtenir une prédiction fiable en utilisant un plus petit ensemble de pseudo-mots. De la même façon que les travaux de (Marczyk *et al.*, 2020) ont prouvé que nous pouvions réduire la liste aux seuls 16 pseudo-mots ayant une double consonne pour le score DAP des évaluateurs humains, l'idée est ici d'observer les résultats sur la prédiction automatique.

La structure des pseudo-mots utilisées pour la réduction est décrite dans la sous-section 3.1. Pour le score automatique, les résultats individuels de chaque pseudo-mot sont conservés, sans nouvel entraînement du modèle. Le score final de chaque patient est alors le score des sous-ensembles des mots de la liste réduite tandis que son score de référence est toujours la moyenne des 52 scores perceptifs. Les résultats sur la liste réduite sont présentés dans le tableau 3.

Au niveau de la corrélation et de l'erreur pour les ensembles avec 16 mots, nous ne voyons pas un gros changement par rapport aux résultats obtenus avec 52 mots. Cet aspect corrobore le fait trouvé dans (Marczyk *et al.*, 2020), qu'il est possible d'obtenir une prédiction fiable en utilisant un ensemble de mots significativement plus petit, ici, en utilisant des mesures automatiques.

Pour les autres sous-ensembles, les résultats sont similaires, néanmoins, nous voyons que dans les sous-ensembles de 10 pseudo-mots, il y a un plus grand écart, au niveau de la corrélation et de l'erreur, entre les mots avec et sans double consonne. L'ensemble des 10 pseudo-mots sans double consonne étant le moins bon. Cela corrobore le fait que les mots avec double consonne sont plus pertinents dans l'obtention d'une mesure d'intelligibilité automatique. Ceci est aussi illustré avec les résultats de l'ensemble final de 5 pseudo-mots avec les deux occurrences des doubles-consonnes, que n'ont pas montré un gros changement au niveau de l'erreur par rapport à, par exemple, l'ensemble de 26 pseudo-mots sans double consonne.

TABLE 3 – Comparaison entre les scores précédemment obtenus sur la liste de pseudos-mots complète et ceux des listes réduites. L'acronyme *d.c.* signifie double consonne.

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

4.2 Discussion

Les résultats de ce travail ont montré qu'il est possible d'obtenir une forte corrélation entre une évaluation perceptive et un score automatique utilisant des réseaux de neurones profonds avec attention. De plus, les valeurs de corrélation ont considérablement augmenté et l'erreur quadratique moyenne obtenue avec notre approche est diminuée de moitié en comparaison avec la méthode précédente (Algorithme Wagner-Fischer). Des mesures automatiques fiables peuvent donc être calculées pour prédire l'intelligibilité.

De plus, l'analyse du nombre de pseudo-mots à faire prononcer par un locuteur a montré qu'il est possible d'enlever des mots et ainsi gagner du temps de passation, tout en continuant à obtenir de bonnes prédictions. L'utilisation de mots avec des consonnes doubles s'étant avérée cruciale dans cette partie. Comme dans l'étude au niveau perceptuel (Marczyk *et al.*, 2020), nous croyons que le contenu phonétique plus important et les co-articulations de consonnes clés sont à l'origine des résultats obtenus avec les sous-ensembles des doubles consonnes.

Tout en conservant peu de pseudo-mots, les résultats obtenus avec le sous-ensemble contenant uniquement les deux occurrences à double consonne sont encourageants : ceci laisse la possibilité de réduire encore plus la quantité de pseudo-mots sans trop affecter les valeurs de base pour la corrélation et l'erreur. Cette réduction de la liste à faire prononcer au patient reste très pertinente dans le contexte clinique, afin de concevoir des batteries d'examens plus courtes pour les patients. La création d'une liste réduite de mots, contenant un contenu phonétique riche et diversifié pouvant fonctionner à la fois au niveau perceptif et automatique, est une piste intéressante pour nos futurs travaux.

5 Conclusions

Cet article a étudié une nouvelle façon d'effectuer une prédiction automatique de l'intelligibilité de la parole : en utilisant des réseaux de neurones profonds avec mécanisme d'attention. Notre approche est appliquée à une tâche de lecture de 52 pseudo-mots par locuteur dans le contexte de cancer des voies aérodigestives supérieures. Une forte corrélation ($\rho = 0,87$) est obtenue entre la moyenne des scores des mots de chaque patient. L'erreur est également faible, montrant que nous pouvons obtenir une prédiction automatique fiable de l'intelligibilité de la parole en utilisant notre méthodologie. Nos résultats sont meilleurs que les précédentes études, aussi bien sur la valeur de corrélation, que sur l'erreur quadratique moyenne. En effet, cette dernière est même divisée par deux. Même en réduisant l'ensemble des pseudo-mots utilisés, nous conservons une très bonne corrélation avec le score perceptif de référence. Ces résultats sont encourageants en vue d'une utilisation clinique de notre méthode, suggérant que nous pouvons concevoir une passation (batterie d'examens) beaucoup plus courte afin d'évaluer l'intelligibilité de la parole.

Remerciements

Ce projet a reçu un financement du programme de recherche et d'innovation Horizon 2020 de l'Union européenne dans le cadre de la convention de subvention Marie Skłodowska-Curie n°766287.

Références

BALAGUER M., POMMÉE T., FARINAS J., PINQUIER J., WOISARD V. & SPEYER R. (2019). Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis : Systematic review. *Journal of the Sciences and Specialities of Head and Neck*.

CHRISTENSEN H., CUNNINGHAM S., FOX C., GREEN P. & HAIN T. (2012). A comparative study of adaptive, automatic recognition of disordered speech. *Proceedings of Interspeech*.

DONG L., XU S. & XU B. (2018). Speech-transformer : A no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5884–5888.

FREDOUILLE C., GHIO A., LAARIDH I., LALAIN M. & WOISARD V. (2019). Acoustic-phonetic decoding for speech intelligibility evaluation in the context of head and neck cancers. *International Congress of Phonetic Sciences (ICPhS)*.

GHIO A., LALAIN M., GIUSTI L., POUCHOULIN G., ROBERT D., REBOURG M., FREDOUILLE C., LAARIDH I. & WOISARD V. (2018). Une mesure d'intelligibilité par décodage acoustico-phonétique de pseudo-mots dans le cas de parole atypique. *XXXIIe Journées d'Études sur la Parole*.

MARCZYK A., GHIO A., LALAIN M., REBOURG M., FREDOUILLE C. & WOISARD V. (2020). Have a cake and eat it too : Assessing discrimination performance of an intelligibility index obtained from a reduced sample size. *12th Conference on Language Resources and Evaluation*.

MIDDAG C. (2012). *Automatic analysis of pathological speech*. Doctoral Dissertation : Ghent University, Department of Electronics and information systems, Ghent, Belgium.

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. *Proceedings of Interspeech*.

VASWANI A., PARMAR N. S. N., USZKOREIT J., JONES L., GOMEZ A. N., ŁUKASZ KAISER & POLOSUKHIN I. (2017). Attention is all you need. *31st Conference on Neural Information Processing System*.

WOISARD V., ASTÉSANO C., BALAGUER M., FARINAS J., FREDOUILLE C., GAILLARD P., GHIO A., GIUSTI L., LAARIDH I., LALAIN M., LEPAGE B., MAUCLAIR J., NOCAUDIE O., PINQUIER J., POUCHOULIN G., PUECH M., ROBERT D. & ROGER V. (2020). C2SI corpus : a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers. *Language Resources and Evaluation*.