Les représentations de locuteurs pour prédire l'intelligibilité de la parole lors de conversations médicales

Sebastião Quintas, Mathieu Balaguer, Julie Mauclair, Virginie Woisard, Julien Pinquier

Les représentations de locuteurs pour prédire l'intelligibilité de la parole lors de conversations médicales

Sebastião Quintas[1] Mathieu Balaguer[1, 2] Julie Mauclair[1] Virginie Woisard[2, 3]
Julien Pinquier[1]
(1) IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
(2) Hôpital Larrey, Toulouse, France
(3) Laboratoire de NeuroPsychoLinguistique, UR 4156, Université de Toulouse, Toulouse, France
sebastiao.quintas@irit.fr

RÉSUMÉ

Dans le contexte des troubles de la parole, l'une des tâches du thérapeute est de définir l'intelligibilité de la parole du patient. Les systèmes automatiques peuvent aider dans cette tâche, mais dans la plupart des cas, ils sont entraînés dans des environnements spécifiques et contrôlés, avec des conditions propres qui ne reflètent pas un environnement médical. Dans cet article, nous développons un système automatique qui prédit l'intelligibilité de la parole à partir de données provenant de patients ayant un cancer de la tête et du cou obtenues dans des conditions cliniques. Ce système repose sur des représentations de locuteurs entraînées selon une méthodologie multi-tâches pour prédire simultanément l'intelligibilité de la parole et la sévérité des troubles de la parole. Il atteint une corrélation allant jusqu'à 0,891 pour une tâche de lecture. De plus, il affiche des résultats prometteurs sur de la parole spontanée, qui est une tâche plus écologique mais sous-étudiée et pourtant essentielle pour un déploiement direct d'un système automatique dans un environnement hospitalier.

ABSTRACT

Speaker Embeddings to Predict Speech Intelligibility in Medical Conversations

In the context of speech disorders, one of the therapist task is to asses the speech intelligibility of a patient. Automatic systems can help in that task but in most cases, they are trained in specific controlled environments with clean conditions that do not reflect a healthcare environment. In this paper, we develop an automatic system that predict speech intelligibility on head and neck cancer data obtained in clinical conditions. This system relies on speaker embeddings trained using a multi-task methodology to simultaneous predict speech intelligibility and speech disorder severity. It achieves a correlation up to 0.891 on a reading task. Morever, it display promosing results on spontaneous speech, which is a more ecologic task yet understudied but nevertheless essential for a direct deployment in a hospital setting.

MOTS-CLÉS : Intelligibilité de la parole, traitement automatique de la parole, représentations de locuteur, cancer de la tête et du cou, parole spontanée.

KEYWORDS: speech intelligibility, automatic speech processing, speaker embeddings, head and neck cancer, spontaneous speech.

1 Introduction

Une altération fonctionnelle au niveau de la communication est généralement présente dès lors qu'un traitement intervient pour des maladies qui affectent les voies aérodigestives supérieures (VADS), telles que le cancer de la tête et du cou (HNC) et les maladies neurodégénératives responsables de dysarthries. Étant donné que des répercussions fonctionnelles majeures sur les VADS sont susceptibles de survenir, une perte d'intelligibilité de la parole est souvent observée, impactant la qualité de vie du patient (de Graeff *et al.*, 2000). En raison du temps nécessaire à la mise en œuvre progressive du post-traitement et de sa durée, un diagnostic précoce est pertinent. Ce diagnostic ainsi que le suivi des troubles sont couramment basés sur une évaluation perceptive de l'intelligibilité de la parole.

Dans les mesures cliniques perceptives, il existe en plus de l'intelligibilité, la sévérité des troubles de la parole qui peut être considérée comme une mesure plus globale incluant la première. Malgré le fait qu'elles servent à deux objectifs différents, ces deux mesures partagent des corrélations élevées : l'une évalue la qualité de parole à un bas niveau, acoustico-phonétique (intelligibilité) et l'autre évalue le degré d'impact du trouble de la parole de façon plus globale sur la communication fonctionnelle (sévérité). De plus, ces mesures sont connues pour être hautement variables, biaisées et subjectives, car leurs évaluations peuvent être conditionnées par la connaissance préalable de la tâche à réaliser (par exemple, la lecture de textes), des évaluations antérieures ou encore une connaissance *a priori* des patients (Fex, 1992). Une approche automatique est alors une alternative pouvant favoriser des prédictions plus fiables et plus objectives.

Les approches pour la prédiction automatique de l'intelligibilité vont de scores basés sur les performances de reconnaissance automatique de la parole (Christensen *et al.*, 2012; Fontan *et al.*, 2017) à des techniques de traitement du signal plus traditionnelles ou à des méthodologies d'apprentissage automatique (Quintas *et al.*, 2022; Bin *et al.*, 2019). Le paradigme de l'embedding de locuteurs, où les énoncés de parole sont représentés par des vecteurs de dimension fixe ayant des propriétés discriminantes entre les locuteurs, a montré des apports intéressants sur des tâches distinctes telles que l'intelligibilité de la parole (Laaridh *et al.*, 2018; Quintas *et al.*, 2020), mais aussi sur l'évaluation générale de la parole pathologique (Zargarbashi & Babaali, 2019; Codosero *et al.*, 2019).

Alors que les travaux récents sur la prédiction automatique de l'intelligibilité de la parole affichent des résultats prometteurs, la majorité des systèmes sont testés sur des données qui ne reproduisent que difficilement les conditions réelles d'un hôpital. Les salles d'enregistrement sont traitées acoustiquement, en utilisant toujours le même microphone, une distance de microphone prédéfinie et des tâches de parole prédéfinies(Clapham *et al.*, 2012; Woisard *et al.*, 2020). Étant donné que l'objectif final de ces systèmes est de fournir des estimations d'intelligibilité plus robustes et écologiques, il devient essentiel de les évaluer sur différents ensembles de données et scénarii cliniques lorsqu'on envisage leur mise en œuvre directe.

De plus, les tâches de parole généralement utilisées pour les évaluations perceptives ou automatiques (Fredouille *et al.*, 2019; Quintas *et al.*, 2020) de la parole sont habituellement la lecture de textes et les pseudo-mots. Les tâches impliquant la parole spontanée, elles, n'ont pas encore été suffisamment étudiées dans le domaine des évaluations automatiques (Balaguer *et al.*, 2019b). Une évaluation automatique sur de la parole spontanée dans des conditions cliniques réelles se rapproche pourtant grandement de l'environnement dans lequel ce type de systèmes serait déployé, tout en utilisant des données représentant étroitement la capacité de communication réelle d'un locuteur.

Par conséquent, dans ce travail, nous menons des expériences sur un système de prédiction d'in-

telligibilité adapté à des données hospitalières plus écologiques. Ainsi, nous avons l'intention de : (i) Analyser la fiabilité d'un système de prédiction de l'intelligibilité/sévérité basé sur des données enregistrées dans des conditions cliniques réelles, et (ii) Évaluer la fiabilité du même système lors de la prédiction de l'intelligibilité/sévérité basée sur des segments de parole spontanée, obtenus à partir d'entretiens patient-soignant.

Le reste de cet article est organisé comme suit. La section 2 décrit notre système et notre méthodologie globale. La section 3 présente notre corpus, nos expériences et nos résultats. Enfin, les sections 4 et 5 proposent une discussion et nos conclusions et perspectives, respectivement.

2 Méthodologie

De manière similaire à (Quintas *et al.*, 2020), le système automatique de prédiction de l'intelligibilité utilise le paradigme de représentation de locuteurs et un réseau de neurones. Dans cet article, le système [1] est adapté pour prédire deux mesures perceptives dans un cadre multi-tâches : l'intelligibilité de la parole (INT), définie comme le degré selon lequel le message du locuteur peut être compris par un auditeur, et la sévérité des troubles de la parole (SEV) définie comme le degré d'altération de l'intelligibilité associé à d'autres variables du signal de parole telles que la qualité d'émission du code acoustico-phonétique, la vitesse de parole et d'autres paramètres temporels ou prosodiques pertinents (Balaguer *et al.*, 2019a). Ces deux mesures, bien que partageant une corrélation élevée et un certain degré de similarité, servent à des fins distinctes. Alors que l'intelligibilité évalue directement la qualité de parole d'un patient donné au niveau acoustico-phonétique, la sévérité des troubles de la parole sert de score global de la maladie qui encapsule différents aspects de la communication verbale.

2.1 Représentations de locuteurs

Les représentations de locuteurs sont des représentations de longueur fixe généralement utilisées dans la vérification des locuteurs, la segmentation en locuteurs et la reconnaissance automatique de la parole. Récemment, ils ont montré une capacité à transmettre des attributs du locuteur permettant la détection des troubles affectant la parole (Codosero *et al.*, 2019). Depuis, nous observons une utilisation croissante de ces représentations pour l'évaluation automatique de la parole pathologique. De plus, étant donné la bonne performance de ce paradigme sur des tâches qui traitent généralement d'une parole plus spontanée (par exemple, la segmentation en locuteurs dans des contextes conversationnels) (Larcher *et al.*, 2021), nous émettons l'hypothèse qu'une approche basée sur les représentations pourrait mieux aider à prédire l'intelligibilité de la parole dans ce même contexte, par opposition à la parole lue, généralement utilisée dans les évaluations cliniques. Étant donné que les représentations de locuteurs à base de *x-vecteurs* ont surpassé les *i-vecteurs* dans la prédiction d'intelligibilité (Quintas *et al.*, 2020), deux classes de représentations de locuteurs ont été testées dans la présente étude, toutes deux extraites à l'aide du toolkit Speechbrain (Ravanelli *et al.*, 2021).

Les premières classes sont des *x-vecteurs*. Elles sont utilisées dans (Quintas *et al.*, 2020) pour prédire l'intelligibilité de la parole en extrayant les caractéristiques discriminantes entre les locuteurs (Snyder *et al.*, 2018). L'extraction des représentations [2] fonctionne en faisant passer le signal de parole à travers

1. https://gitlab.irit.fr/samova/embedding_intelligibility
2. https://huggingface.co/speechbrain/spkrec-xvect-voxceleb

un bloc de réseaux neuronaux à retard temporel (TDNN) qui opère sur des trames de parole avec un contexte temporel réduit centré sur la trame actuelle. Les couches TDNN ultérieures s'appuient sur le contexte temporel des couches précédentes. Une couche de regroupement statistique agrège toutes les sorties au niveau des trames en une dimension de longueur fixe, qui est ensuite alimentée dans un bloc entièrement connecté. Les *x-vecteurs* sont extraits à partir de la composante affine de la dernière couche entièrement connectée. Le système a été pré-entraîné avec les données voxceleb1 (Nagrani *et al.*, 2017) et voxceleb2 (Chung *et al.*, 2018), puis testé sur l'ensemble de test voxceleb1, atteignant une erreur (EER, (Cheng & Wang, 2004)) de 3,2%.

Ensuite, les représentations de locuteurs Ecapa TDNN [3], plus récents que les *x-vecteurs*, ont été expérimentés. Ces représentations de longueur fixe s'appuient sur le concept des *x-vecteurs*, avec cependant plusieurs améliorations qui suggèrent une meilleure performance en vérification des locuteurs par rapport à d'autres représentations (Desplanques *et al.*, 2020). Nous supposons que ces améliorations permettent au réseau de se concentrer davantage sur les caractéristiques du locuteur qui ne s'activent pas aux mêmes instants, par exemple les propriétés spécifiques du locuteur sur les voyelles par rapport aux propriétés spécifiques du locuteur sur les consonnes. Nous émettons l'hypothèse que ces améliorations pourraient fournir un embedding de locuteur plus robuste pour l'évaluation de la parole pathologique, et par conséquent surpasser les *x-vecteurs* précédemment utilisés. De manière similaire à l'extracteur précédemment introduit, le système Ecapa TDNN a été pré-entraîné en utilisant les données voxceleb1 et voxceleb2, puis testé sur l'ensemble de test voxceleb1, atteignant une EER de seulement 0,8%.

2.2 Réseau neuronal

La figure 1 présente un diagramme de notre réseau neuronal. Le réseau reçoit en entrée les représentations de locuteurs qui, selon le type, ont des dimensions fixes distinctes (512 pour les *x-vecteurs* et 192 pour l'Ecapa TDNN). De plus, le signal passe à travers deux couches de dimensions fixes, puis enfin les deux couches multi-tâches qui prédisent les deux mesures perceptives différentes. Le système est optimisé à l'aide d'une fonction de perte de l'erreur quadratique moyenne (MSE) et d'un algorithme d'optimisation Adam. Afin de prédire deux mesures distinctes, la fonction de perte prend en compte ces deux mesures avec des contributions égales, ce qui signifie un poids de 50 % pour l'intelligibilité et 50 % pour la sévérité. En raison des corrélations élevées généralement observées entre ces deux mesures, nous émettons l'hypothèse que leur apprentissage conjoint conduira à une estimation de l'intelligibilité de la parole meilleure et plus robuste.

2.3 Entraînement et validation

Un système pour chaque type de représentations a ainsi été entraîné sur le corpus C2SI (Woisard *et al.*, 2020). Le corpus comprend une variété de patients souffrant de cancer de la tête et du cou avec des localisations tumorales initiales différentes, ainsi que des locuteurs sains. Les deux systèmes ont été entraînés et validés en utilisant la tâche de lecture de texte segmentée. Un schéma d'augmentation des données, similaire à celui de (Quintas *et al.*, 2020), basé sur une distorsion temporelle (Vachhani *et al.*, 2018; Ko *et al.*, 2015) qui préserve le timbre et l'enveloppe spectrale, a été implémenté pendant l'entraînement. Un total de 98 locuteurs a été utilisé pour l'entraînement et un sous-ensemble de 10

3. https://huggingface.co/speechbrain/spkrec-ecapa-voxceleb

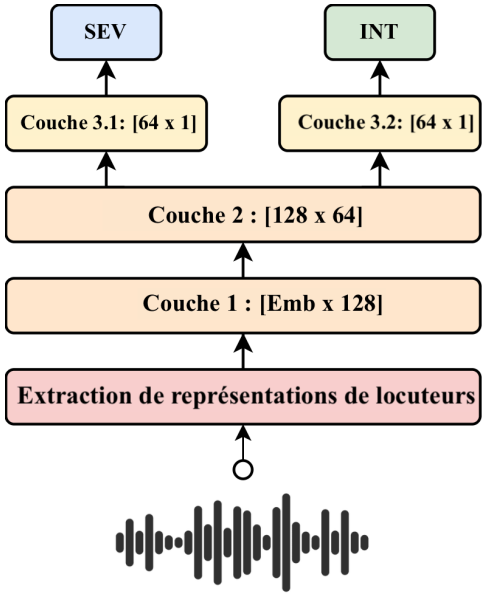

FIGURE 1 – Diagramme schématique du réseau neuronal proposé et de l'approche d'apprentissage multi-tâches correspondante. La mention "Emb" indique la taille de la représentation de locuteurs.

locuteurs avec des degrés d'intelligibilité variables a été utilisé comme ensemble de validation. Un batchsize de 8, un taux d'apprentissage de 0,001 et une fonction de perte (loss) de 0,2 ont été utilisés sur 20 epochs.

3 Expériences et Résultats

3.1 Corpus SpeeCOmco

Notre corpus de parole et de communication en oncologie (SpeeCOmco) est un ensemble de 27 patients présentant des degrés d'intelligibilité variables ayant enregistré différentes tâches dans des conditions cliniques réelles (Balaguer, 2021). Dans la population du corpus, l'âge moyen est de 66,3 ans (min. 38 ans, max. 83 ans) avec une représentation féminine de 37%. Les enregistrements ont été réalisés dans des salles de consultation, non traitées acoustiquement, avec l'utilisation d'un micro-casque couramment utilisé en pratique clinique et la présence d'un certain niveau de bruit de fond. Les enregistrements ayant eu lieu dans un environnement hospitalier, plus précisément lors de rendez-vous cliniques en orthophonie, les conditions d'enregistrement imitent exactement les conditions dans lesquelles le présent système serait déployé. Tous les patients sont des locuteurs natifs du français.

Pour cet ensemble de patients, l'intelligibilité moyenne et la sévérité des troubles de la parole ont été calculées sur la base d'une évaluation perceptive indépendante de six professionnels de la santé. Chaque locuteur a reçu un score entre 0 et 10, plus la valeur est petite, moins la parole est intelligible. La même échelle est utilisée pour la sévérité. Le coefficient de corrélation intraclasse (CCI) a été

calculé pour évaluer la fiabilité inter-juges. Un CCI de 0,816 a été obtenu pour les six juges lors de l'évaluation de l'intelligibilité de la parole parmi tous les patients et un CCI de 0,852 a été obtenu pour la sévérité, montrant un bon niveau d'accord entre les experts.

Plusieurs tâches, classiques dans l'évaluation de la parole pathologique, ont été utilisées :

1. **Lecture de texte (LEC)**. Les locuteurs ont été invités à lire le premier paragraphe de « La chèvre de M. Seguin », un conte d'Alphonse Daudet choisi car il est assez long pour inclure presque tous les phonèmes français. Ce passage est également bien connu et largement utilisé en phonétique clinique française (Ghio *et al.*, 2012).
2. **Phrases avec semi-voyelles (PHR)**. Les locuteurs lisent deux phrases contenant les semi-voyelles françaises [w] et [U], absentes du texte LEC.
3. **Inflexion consonantique (CSN)**. Les locuteurs ont été invités à lire 17 phrases sous la forme de *« Le sac eu**C**eu convient »*, où le ***C*** est remplacé à chaque phrase par une consonne différente.
4. **Pseudo-mots (DAP)**. Chaque locuteur enregistre un ensemble de 52 pseudo-mots, inexistants dans la langue française (Lalain *et al.*, 2020). Chaque pseudo-mot a été généré automatiquement de manière à respecter les règles phonotactiques et orthographiques du français.
5. **Parole spontanée (SPO)**. Dans cette tâche, l'échantillon audio provient d'un entretien entre un orthophoniste et le locuteur. La conversation porte sur la communication quotidienne et les limitations perçues par le locuteur. Les segments de parole spontanée sont obtenus grâce à un détecteur d'activité vocale (VAD). Les segments de moins de 3s et de plus de 10s ont été écartés afin de minimiser le nombre d'artefacts capturés. Les segments avec une trop forte présence de la voix du thérapeute ont également été supprimés. La durée de l'entretien peut ici beaucoup différer entre les locuteurs (ainsi que le nombre de fichiers associés : de 8 à 56 dans ce corpus).

3.2 Tests et résultats

Les 27 patients et les cinq tâches de parole enregistrées ont été évalués à l'aide des deux systèmes décrits précédemment, l'un utilisant les *x-vecteurs* et l'autre les représentations de locuteurs Ecapa TDNN. À l'exception de la tâche de parole spontanée, dont la prédiction d'intelligibilité correspond à la moyenne des fichiers segmentés en parole (via une VAD mentionnée précédemment), les tâches ont été analysées sur un seul fichier audio par locuteur. Le tableau 1 illustre la corrélation de Spearman (ρ) pour les différentes tâches évaluées et le type de représentations, et les valeurs de l'erreur quadratique moyenne (RMSE) sur l'intelligibilité de la parole (ainsi que sur la prédiction de la sévérité des troubles de la parole). Les corrélations sont élevées ($\rho > 0{,}82$) sur quatre des cinq tâches lors de l'utilisation des *x-vecteurs*. De même, les erreurs sont faibles (RMSE < 1,5) sur trois des tâches. La figure 2 affiche un graphique des prédictions associées à la tâche de parole spontanée.

4 Discussion

Les résultats ont montré des valeurs de corrélation et d'erreur encourageantes. De plus, étant donné que l'objectif final de cet article est de valider la mise en œuvre d'un système de prédiction de l'intelligibilité et de sévérité sur le plan clinique, une combinaison de corrélation élevée et d'erreurs

TABLE 1 – Valeurs de corrélation et d'erreur obtenues lors du test du système sur les différentes tâches du corpus SpeeCOmco. Les valeurs en gras marquent les tâches avec les meilleurs résultats. Toutes les corrélations ont atteint une valeur $p < 0{,}05$, les rendant statistiquement significatives.

Mesures perceptives		**Sévérité**				**Intelligibilité**			
Représentations		**Ecapa TDNN**		***X-vecteurs***		**Ecapa TDNN**		***X-vecteurs***	
Métriques d'évaluation		ρ	RMSE	ρ	RMSE	ρ	RMSE	ρ	RMSE
Tâches de parole	**LEC**	0.783	1.873	**0.866**	**1.384**	0.826	2.070	**0.891**	**1.322**
	PHR	0.784	1.782	0.805	1.772	0.807	1.854	**0.842**	**1.460**
	CSN	0.643	2.060	0.861	2.124	0.673	2.147	0.859	1.971
	DAP	0.296	2.673	0.724	2.219	0.371	2.805	0.731	1.881
	SPO	0.657	2.124	0.818	1.820	0.695	2.252	**0.828**	**1.468**

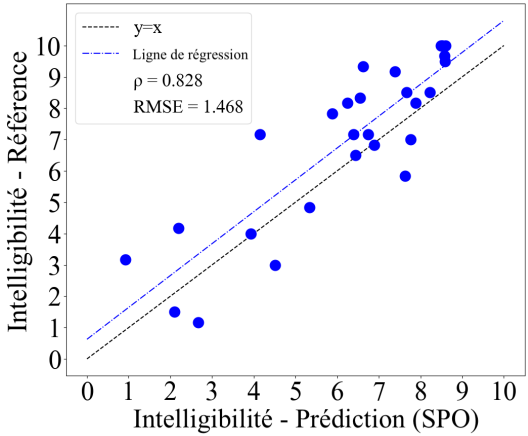

FIGURE 2 – Résultats de la prédiction de l'intelligibilité sur le corpus SpeeCOmco en utilisant la tâche SPO (*x-vecteurs*).

faibles doit être envisagée. Malgré les résultats prometteurs des représentations ECAPA pour la vérification des locuteurs (Desplanques *et al.*, 2020), dans notre contexte spécifique, les *x-vecteurs* les surpassent dans toutes les tâches de parole, tant pour l'intelligibilité que pour la sévérité. Cet aspect valide non seulement l'utilisation des *x-vecteurs* pour la parole pathologique, mais montre également que tous les types de représentations de locuteurs ne conviennent pas à ce type d'analyse. Une étude comparative approfondie sur les différentes représentations de locuteurs pour ce type d'évaluation ainsi que la recherche d'une meilleure métrique pour analyser leur performance sur la parole pathologique est une piste intéressante pour les travaux futurs.

Les *x-vecteurs* ont obtenus des résultats intéressants et fiables, à l'exception de la tâche DAP. D'une part, ceci peut être expliqué par la qualité de ces enregistrements : artefacts de bruit entre les pseudo-mots. D'autre part, aucune prosodie ni coarticulation entre les pseudo-mots ne sont présentes. Le système ayant été entraîné sur la tâche de lecture de texte, bien évidemment, il se comporte mieux

sur cette même tâche LEC. Cependant, la tâche PHR a également obtenue des résultats fiables, bien que les fichiers audio soient beaucoup plus courts que ceux de la tâche de lecture. Enfin, les résultats sur la parole spontanée ont présenté une corrélation forte et une faible erreur (comparable aux autres tâches). L'évaluation de ce type de parole devient très pertinente en raison du fait qu'il s'agit d'un médium peu exploré (Balaguer *et al.*, 2019b), offrant une vision écologique de la capacité réelle du patient à communiquer. Cette évaluation automatique sur la parole spontanée comble un vide dans la littérature concernant le test des approches automatiques sur ce type de parole enregistrée, et peut être considérée comme une pierre angulaire vers des prédictions d'intelligibilité plus pertinentes, plus écologiques et plus fiables.

5 Conclusions et perspectives

Cet article a examiné la fiabilité d'un prédicteur automatique de l'intelligibilité de la parole basé sur des représentations de locuteurs dans des conditions écologiques (consultations cliniques à l'hôpital). Différentes évaluations ont été réalisées dans une méthodologie d'apprentissage multi-tâches. Les résultats ont suggéré une bonne capacité de généralisation, illustrée par des corrélations allant jusqu'à 0,891 et des erreurs de 1,322. Les métriques obtenues sur la tâche de parole spontanée sont non seulement comparables à celles des autres tâches, mais ouvrent également la possibilité d'une utilisation plus large des évaluations automatiques, un sujet actuellement sous-exploré.

Même si l'intelligibilité de la parole est la principale mesure subjective à analyser et à prédire ici, les résultats sont similaires sur la sévérité des troubles de la parole. Cet aspect montre que le paradigme multi-tâches peut être efficace pour ce type de mesure, et permet d'apprendre d'autres mesures perceptives, telles que la prosodie, la résonance et les distorsions phonémiques. De plus, une mesure d'intelligibilité qui peut être considérée comme une combinaison de ces autres paramètres (de Bodt *et al.*, 2002) peut être plus interprétable, avec une valeur ajoutée dans un environnement clinique.

Le système développé a été entraîné sur une tâche de lecture. Malgré la bonne généralisation du système sur une variété de nouveaux patients et de tâches de parole, un entraînement supplémentaire sur d'autres tâches (comme de la parole spontanée), ainsi que sur d'autres langues et maladies (comme la maladie de Parkinson, la sclérose latérale amyotrophique, etc.) pourrait non seulement augmenter les performances, mais aussi rendre le système encore plus robuste. Le développement d'un modèle d'intelligibilité automatique multi-pathologies est une perspective intéressante pour les travaux futurs. Cependant, il convient de le concevoir avec soin, car une solution fonctionnelle pour l'intelligibilité de la parole dans les cancers de la tête et du cou peut ne pas nécessairement correspondre à la meilleure approche pour les maladies neurologiques. Cela est principalement dû au type de problèmes affectant la parole qui diffèrent grandement entre les deux ensembles de maladies, rendant les systèmes conçus difficilement transposables à toutes pathologies.

Au vu de la capacité de généralisation du système proposé sur différents types de données hospitalières dans le contexte des cancers de la tête et du cou, les travaux futurs prévoient la mise en œuvre directe du présent système en clinique, grâce à une application mobile qui sera utilisée par les thérapeutes.

6 Remerciements

Ce projet a reçu un financement du programme de recherche et d'innovation Horizon 2020 de l'Union européenne dans le cadre de l'accord de subvention Marie Skłodowska-Curie No 766287.

Références

BALAGUER M. (2021). *Mesure de l'altération de la communication par analyses automatiques de la parole spontanée après traitement d'un cancer oral ou oropharyngé*. Thèse de doctorat, Université Paul Sabatier - Toulouse III.

BALAGUER M., BOISGUÉRIN A., GALTIER A., GAILLARD N., PUECH M. & WOISARD V. (2019a). Assessment of impairment of intelligibility and of speech signal after oral cavity and oropharynx cancer. *European Annals of Otorhinolaryngology, Head and Neck Diseases*, **136(5)**, 355–359.

BALAGUER M., POMMÉE T., FARINAS J., PINQUIER J., WOISARD V. & SPEYER R. (2019b). Effects of oral and oropharyngeal cancer on speech intelligibility using acoustic analysis : Systematic review. *Journal of the Sciences and Specialities of Head and Neck*, **42(1)**, 111–130.

BIN L., KELLEY M. C., AALTO D. & TUCKER B. V. (2019). Automatic speech intelligibility scoring of head and neck cancer patients with deep neural networks. *International Congress of Phonetic Sciences (ICPHs')*.

CHENG J.-M. & WANG H.-C. (2004). A method of estimating the equal error rate for automatic speaker verification. *Proceedings of ISCSLP*.

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

CHUNG J. S., NAGRANI A. & ZISSERMAN A. (2018). Voxceleb2 : Deep speaker recognition. *Proceedings of Interspeech*.

CLAPHAM R., VAN DER MOLEN L., VAN SON R., VAN DEN BREKEL M. & HILGERS F. (2012). Nki-ccrt corpus - speech intelligibility before and after advanced head and neck cancer treated with concomitant chemoradiotherapy. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*.

CODOSERO J. M. P., ESPINOZA-CUADROS F., ANTÓN-MARTÍN J., BARBERO-ALVAREZ M. A. & GÓMEZ L. A. H. (2019). Modeling obstructive sleep apnea voices using deep neural network embeddings and domain-adversarial training. *IEEE Journal of Selected Topics in Signal Processing*, **14(2)**, 240–250.

DE BODT M., HUICI M. E. & HEYNING P. V. D. (2002). Intelligibility as a linear combination of dimensions in dysarthric speech. *Journal of Communication Disorders*, **35(3)**, 283–292.

DE GRAEFF A., DE LEEUW R. J., ROS W. J., HORDIJK G.-J., BLIJHAM G. H. & WINNUBST J. A. (2000). Long-term quality of life of patients with head and neck cancer. *The Laryngoscope, Volume 110, Issue 1*, p. 98–106.

DESPLANQUES B., THIENPONDT J. & DEMUYNCK K. (2020). ECAPA-TDNN : Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. *Proceedings of Interspeech*.

FEX S. (1992). Perceptual evaluation. *Journal of Voice*, **6**(2), 155–158.

FONTAN L., FERRANÉ I., FARINAS J., PINQUIER J., TARDIEU J. & MAGNEN C. (2017). Automatic speech recognition predicts speech intelligibility and comprehension for listeners with simulated age-related hearing loss. *Journal of Speech, Language and Hearing Research, Volume 50(1)*, **60(9)**, 2394–2405.

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)*, p. 3051–3055.

GHIO A., POUCHOULIN G., TESTON B., PINTO S., FREDOUILLE C. & ET AL (2012). How to manage sound, physiological and clinical data of 2500 dysphonic and dysarthric speakers ? *Speech Communication*, **54**, 664–679.

KO T., PEDDINTI V., POVEY D. & KHUDANPUR S. (2015). Audio augmentation for speech recognition. *Proceedings of Interspeech*.

LAARIDH I., FREDOUILLE C., GHIO A., LALAIN M. & WOISARD V. (2018). Automatic evaluation of speech intelligibility based on i-vectors in the context of head and neck cancers. *Proceedings of Interspeech*.

LALAIN M., GHIO A., GIUSTI L., ROBERT D., FREDOUILLE C. & WOISARD V. (2020). Design and development of a speech intelligibility test based on pseudowords in french : Why and how ? *Journal of Speech, Language and Hearing Research*, **63(7)**, 2070–2083.

LARCHER A., MEHRISH A., TAHON M., MEIGNIER S., CARRIVE J., DOUKHAN D., GALIBERT O. & EVANS N. (2021). Speaker embeddings for diarization of broadcast data in the allies challenge. *Proceedings of ICASSP*.

NAGRANI A., CHUNG J. S. & ZISSERMAN A. (2017). Voxceleb : A largescale speaker identification dataset. *Proceedings of Interspeech*.

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*.

QUINTAS S., MAUCLAIR J., WOISARD V. & PINQUIER J. (2022). Automatic assessment of speech intelligibility using consonant similarity for head and neck cancer. *Proceedings of Interspeech*.

RAVANELLI M., PARCOLLET T., PLANTINGA P., ROUHE A., CORNELL S., LUGOSCH L., SUBAKAN C., DAWALATABAD N., HEBA A., ZHONG J., CHOU J.-C., YEH S.-L., FU S.-W., LIAO C.-F., RASTORGUEVA E., GRONDIN F., ARIS W., NA H., GAO Y., MORI R. D. & BENGIO Y. (2021). SpeechBrain : A general-purpose speech toolkit. *arXiv :2106.04624*.

SNYDER D., GARCIA-ROMERO D., SELL G., POVEY D. & KHUDANPUR S. (2018). X-vectors : Robust dnn embedings for speaker recognition. *Proceedings of ICASSP*.

VACHHANI B., BHAT C. & KOPPARAPU S. K. (2018). Data augmentation using healthy speech for dysarthric speech recognition. *Proceedings of Interspeech*.

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*, **55**, 173–190.

ZARGARBASHI S. & BABAALI B. (2019). A multi-modal feature embedding approach to diagnose alzheimer disease from spoken language. *arXiv :1910.00330*.