XXXIVe Journées d'Études sur la Parole – JEP 2022
13-17 June 2022, Île de Noirmoutier, France

Création d'une mesure entropique de la parole pour évaluer l'intelligibilité de patients atteints de cancers des voies aérodigestives supérieures

Vincent Roger[1] Jérôme Farinas[1] Virginie Woisard[2,3] Julien Pinquier[1]
(1) IRIT, Université de Toulouse, CNRS, Toulouse INP, UT3, Toulouse, France
(2) Institut Universitaire du Cancer de Toulouse, Centre Hospitalo Universitaire de Toulouse, Service ORL de l'Hôpital Larrey, Toulouse, France
(3) Laboratoire de NeuroPsychoLinguistique, UR 4156, Université de Toulouse, Toulouse, France
{Vincent.Roger, Jerome.Farinas, Julien.Pinquier}@irit.fr,
woisard.v@chu-toulouse.fr

RÉSUMÉ

L'évaluation de la sévérité de la parole, valeur fortement corrélée à l'intelligibilité de prononciation, est importante pour les cliniciens afin de mesurer l'impact des cancers des voies aérodigestives supérieures. Nous proposons une nouvelle méthode de mesure entropique, fondée sur des représentations de la parole, qui produit un score semblable à l'indice de sévérité. Cette méthode est fondée sur l'Inception Score (généralement utilisé en image pour évaluer la qualité des images générées par les modèles). Sur le corpus cancer C2SI, nous obtenons une corrélation de Spearman de 0,87 avec un indice de sévérité évalué manuellement par un groupe d'experts cliniques. Cette forte corrélation nous permet d'envisager des applications cliniques.

ABSTRACT

Creation of an entropic speech measurement to evaluate the intelligibility of patients with head and neck cancers

The assessment of speech severity, a value that is highly correlated with speech intelligibility, is important for clinicians to measure the impact of head and neck cancers. We propose a new entropic measurement method, based on speech representations, which produces a score similar to the severity index. This method is based on the Inception Score (generally used in imaging to evaluate the quality of images generated by models). On the C2SI cancer corpus, we obtain a Spearman correlation ($\rho = 0.87$) with a severity index manually evaluated by a group of clinical experts. This strong correlation allows us to consider clinical applications.

MOTS-CLÉS : Score entropique, sévérité, intelligibilité, encodeur parole, cancer des voies aérodigestives supérieures.

KEYWORDS: Entropic score, severity, intelligibility, speech embeddings, head and neck cancer.

1 Introduction

Entre 2007 et 2016, le nombre de cancers de la cavité buccale et du pharynx combinés a augmenté, avec une incidence annuelle de 3% aux États-Unis (Ellington, 2020). De nombreuses autres patho-

10.21437/JEP.2022-13

logies peuvent provoquer des troubles de la parole en affectant les mécanismes de production de la parole (Enderby, 2013). Ceci peut réduire la qualité de vie des patients affectés (Walshe & Miller, 2011). Pour quantifier l'impact de ces pathologies sur la production de la parole, il est courant d'utiliser des questionnaires évaluant le degré de handicap ainsi que des scores perceptifs qui représentent l'intelligibilité de la production et l'indice de sévérité du handicap concernant la parole. Les notions d'intelligibilité et de sévérité diffèrent beaucoup au sein des experts de la communauté. Suite à l'étude qui a abouti à un consensus (Pommée *et al.*, 2021), nous utiliserons donc dans ce papier le terme sévérité pour désigner le degré d'altération de la parole. Celui-ci ne mesure pas la qualité du signal audio, mais la capacité des patients à produire des sons intelligibles. Les experts obtiennent ce score en écoutant leurs patients. La familiarité avec les patients et l'habitude d'entendre des voix pathologiques peuvent influencer le score produit par les experts (Landa *et al.*, 2014). Afin d'augmenter la stabilité de la mesure, celle-ci peut être traitée par un groupe d'experts (Woisard & Lepage, 2010). Néanmoins, le recours à un groupe d'experts pour obtenir un score objectif complique le suivi des patients. L'utilisation d'un outil automatique, calculé sur un échantillon de parole enregistré, pourrait fournir une solution objective et libérer du temps pour les experts. Dans cet article, nous proposons une nouvelle méthode automatique pour évaluer un tel score. Nous nous concentrons sur l'indice de sévérité et sur les troubles de la parole des voies aérodigestives supérieures (entre autres, les productions de la cavité buccale et pharyngée).

Dans certains articles récents, nous avons recensé des approches aveugles (approches qui construisent un score en utilisant uniquement les enregistrements des patients, selon (Janbakhshi *et al.*, 2019)) (Fang *et al.*, 2017; Fletcher *et al.*, 2017) et des approches non aveugles (utilisant des patients et des contrôles pour créer leur score, selon (Janbakhshi *et al.*, 2019)) (Laaridh *et al.*, 2017; Janbakhshi *et al.*, 2019).

La méthode que nous proposons se place dans la catégorie « non aveugle » et utilise un score entropique sur un encodeur de la parole. Elle diffère des autres méthodes, parce qu'elle utilise une approche non supervisée : le score ne résulte pas d'une phase d'apprentissage sur le corpus cible, il est construit en utilisant une mesure entropique sur le paramètre dominant du signal (en fonction d'un encodeur donné). Nous avons focalisé nos expériences sur la tâche de lecture du corpus Carcinologic Speech Severity Index (C2SI) (Woisard *et al.*, 2020).

Préliminairement à ce travail, nous avons essayé d'apprendre des modèles avec une régression directe entre le score de sévérité et le signal de parole. Nous avons également utilisé l'apprentissage par transfert (à l'aide du modèle PASE+ (Ravanelli *et al.*, 2020)), mais nous n'avons obtenu que des performances médiocres. L'une des raisons probables à ces résultats est que, dans la base de données C2SI, les valeurs des scores de sévérité ne couvrent pas uniformément toutes les plages de valeurs : il y a moins de représentants avec des scores faibles que de représentants avec des scores élevés. En effet, l'acquisition de résultats pour toutes les plages possibles de valeurs est une tâche difficile, voire impossible. La nécessité de créer des ensembles d'entraînement et de tests pour les tâches de régression implique que moins de données sont disponibles pour l'entraînement.

Pour résoudre ce problème « classique » de manque de données en parole pathologique, nous proposons une nouvelle approche en section 2 : il s'agit de réaliser un score entropique comparable à l'Inception Score (IS) (Salimans *et al.*, 2016). Nous validons notre méthode sur le corpus C2SI en détaillant nos expériences en section 3.

2 Création d'un score entropique de la parole

Dans notre méthode, notre score est une mesure entropique calculée à partir de représentations spectrales, cepstrales, ou apprises par un réseau de neurones. Ceci nous permet d'éviter d'utiliser des données de notre corpus cible, et d'exploiter des corpus externes pour créer des représentations de notre tâche.

Nous proposons des adaptations de ces représentations pour calculer un score entropique semblable à l'Inception Score (IS) (Salimans *et al.*, 2016). Ainsi, nous proposons une formulation d'un score entropique fondé sur les caractéristiques des signaux émis par chaque participant. De cette façon, nous pouvons comparer les scores obtenus par les patients avec ceux des contrôles (personnes considérées « saines »). Ici, les scores obtenus par les contrôles servent de référence et les scores des patients doivent se rapprocher au maximum de ces scores. Cette comparaison peut être vue comme une mesure de l'altération induite par la maladie. Dans les sous-sections qui suivent, nous détaillerons ce score et nous montrerons et analyserons les résultats obtenus avec une telle approche.

2.1 Présentation de l'approche

Dans le domaine de la vision par ordinateur, la communauté scientifique utilise l'IS comme une métrique permettant de voir dans quelle mesure les données générées par un modèle sont proches des données réelles (Salimans *et al.*, 2016). Ce score mesure la qualité des images créées par des modèles génératifs (notamment les réseaux adversaires génératifs). L'IS implique des prédictions du modèle Inception (Szegedy *et al.*, 2016) (d'où le nom IS de l'approche qui peut porter à confusion) et est une combinaison de deux attentes entropiques sur les probabilités sorties du modèle Inception. Il permet d'évaluer les différences entre plusieurs générateurs d'images pour voir dans quelle mesure leurs performances diffèrent du score obtenu par des échantillons réels. Un exemple d'utilisation de cette mesure se trouve sur la figure 1.

FIGURE 1 – Exemple d'utilisation de l'IS pour distinguer des images venant de cifar10 et provenant de cinq autres modèles. Figure reprise et traduite du papier de (Salimans *et al.*, 2016).

Nous adoptons cette approche aux locuteurs de notre corpus cancer. Nous utilisons les signaux audio pour inférer un score comparable à l'indice de sévérité des troubles de la parole. Dans notre méthode, nous faisons les hypothèses suivantes :

— chaque participant est un générateur,
— chaque participant prononce la même phrase $\mathcal{S}$, afin de garantir que les sons produits appartiennent au même domaine et sont par conséquent comparables,
— le groupe des contrôles représente la qualité de la parole à atteindre par les patients.

Compte tenu de ces hypothèses, nous avons calculé un score pour chaque participant. La figure 2 illustre le traitement que nous proposons. L'entrée correspond à une séquence d'échantillons $\boldsymbol{x}$

représentant le signal produit par un locuteur l. La séquence $\boldsymbol{x}$ forme une phrase prédéfinie $\mathcal{S}$.

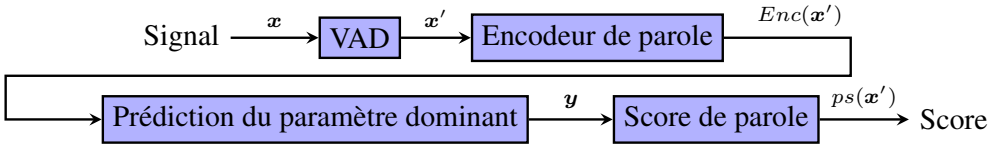

FIGURE 2 – Suite de traitements de notre approche non supervisée.

Notre approche utilise un détecteur d'activité vocale (VAD) et ses implications seront discutées dans la section 2.2.

Les scores obtenus du groupe de contrôle représentent les scores à atteindre (comme les données réelles pour le Score d'Inception) par les patients (les données factices, ou ici les voix dégradées). Plus le score d'un patient est proche des scores de contrôle, moins sa parole a été dégradée. L'objectif est d'obtenir une mesure de la qualité de la parole comparable à l'indice de sévérité des troubles de la parole. En comparaison à l'IS, nous avons fait quelques adaptations : nous utilisons des représentations (apprises par un modèle ou des paramètres acoustiques) au lieu d'un modèle de classification. Dans nos expériences, nous avons utilisé l'encodeur PASE+ (Ravanelli *et al.*, 2020) comme représentation, mais nous avons également essayé des paramètres acoustiques, tels les MFCC et les MEL spectrogrammes. Néanmoins d'autres encodeurs ou représentations du signal peuvent être utilisés, tant les caractéristiques véhiculent le même type d'information (voir la section 2.3 pour une description détaillée).

La sortie des représentations devant être normalisée, nous proposons de transformer ces caractéristiques en probabilités : plus de détails sont disponibles dans la section 2.4. Enfin, nous utilisons ces probabilités (telle une sortie du modèle Inception) pour calculer un score entropique comparable à l'IS (voir la sous-section 2.5).

2.2 Détection d'activité vocale

La première étape de notre traitement consiste à supprimer les longs silences à l'aide d'un VAD. Soit $\boldsymbol{x}'$ les échantillons actifs d'un signal de parole extrait de VAD($\boldsymbol{x}$), VAD étant toute technique qui préserve les silences courts et la voix du locuteur (parole). Pour nos expériences, notre VAD utilise des seuils sur l'énergie et l'aplatissement spectral du percentile 95% inspirés par (Moattar & Homayounpour, 2009). Il prend des décisions pour des fenêtres de 64 ms et les silences de moins de 320 ms sont gardés pour s'assurer que nous ne supprimons que les longs silences. Ce système obtient un F-score supérieur à 97% sur notre corpus.

Notre méthode n'intègre pas les longues latences entre les mots (pouvant être dues à une gêne de la maladie). En effet, pour chaque locuteur, nous mesurons uniquement un score global de la qualité de la production. Même si les longs silences sont une information cruciale pour quantifier l'intelligibilité ou la sévérité, nous les évitons : il s'agit d'une conséquence de notre hypothèse que les locuteurs sont considérés comme des générateurs. Effectivement, avec notre approche, les silences longs représentent une variété de production du générateur qui peut fausser les résultats souhaités. Après avoir extrait le signal actif (parole), nous calculons ses représentations.

2.3 Encodeur de parole

L'encodeur de parole doit décrire le signal $\boldsymbol{x}'$ en une séquence de représentations aux dimensions fixes. Il doit idéalement éviter de coder les bruits environnementaux présents dans le signal et doit se concentrer sur les représentations de la parole (notre normalisation permet d'éviter les perturbations par des bruits constants).

Désignons par Enc l'encodeur de parole ou tout paramètre acoustique. Les représentations produites par Enc doivent réduire la dimension du signal d'entrée. Pour la partie expérimentation, nous avons choisi les MFCC, les Mel spectrogrammes et l'encodeur PASE+ (Ravanelli *et al.*, 2020).

Nous avons choisi PASE+, car il est conçu pour être un encodeur de signaux vocaux générique, adapté à n'importe quelle tâche vocale, avec des trames proches du niveau phonémique. Les auteurs ont entraîné un encodeur à l'aide de plusieurs tâches autosupervisées (telles que la reconstruction du signal, la reconstruction de MFCC, la reconstruction de l'énergie, etc.) et de tâches binaires (Local Info Max et Global Info Max). Le fait de partager le même encodeur pour toutes les tâches oblige cet encodeur à représenter de multiples représentations de la parole. De plus, les auteurs de PASE+ ont entraîné leur modèle en utilisant des mécanismes de débruitage qui répondent à nos besoins.

Nous détaillerons les choix de paramètres vocaux et d'encodeur dans la section 3. Après avoir extrait ces encodages, nous allons transformer ces représentations pour avoir des propriétés semblables au modèle Inception.

2.4 Prédiction du paramètre dominant

Pour garantir un type de sortie semblable à celui du modèle Inception, utilisé pour la métrique IS, nous proposons d'adapter la sortie de l'encodeur en un classifieur de paramètre dominant. Pour calculer cette séquence de probabilités $\boldsymbol{y}$, nous proposons :

$$\boldsymbol{y} = \text{softmax}(\text{abs}(f_{\text{norm}}(Enc(\boldsymbol{x}')))) \tag{1}$$

avec f_{norm} étant une fonction de normalisation établie sur les espaces latents correspondant au signal vocalisé ($Enc(\boldsymbol{x}')$).

Dans nos expériences, nous avons testé plusieurs normalisations : L_1, $L_{\max}$, L_∞. Nous avons appliqué ces normalisations pour chaque dimension de nos différents pas de temps.

Notons $\boldsymbol{X}'_{t,n}$ pour définir la représentation latente d'une fenêtre temporelle ($Enc(\boldsymbol{x}'_t)$) avec n une dimension des paramètres calculés à l'instant t. Ainsi, pour les normalisations, nous avons :

$$f_{\text{norm}}(\boldsymbol{X}'_{t,.}) = \frac{\boldsymbol{X}'_{t,.}}{||\boldsymbol{X}'_{.,.}||_k} \tag{2}$$

avec $k \in \{1, \max, \infty\}$ et $\boldsymbol{X}'$ étant toutes les représentations latentes des fenêtres temporelles vocalisées d'un fichier. Notez l'utilisation du point pour signifier l'utilisation de l'ensemble des trames ou des dimensions de $\boldsymbol{X}$.

Nous avons, en plus de ces normalisations, utilisé le $zscore$. Dans ce cas, nous avons calculé la moyenne et l'écart-type pour chaque fichier audio de lecture.

L'utilisation de la fonction absolue dans l'équation 1 permet aux valeurs fortement négatives d'avoir un impact équivalent aux valeurs positives sur le score entropique. Effectivement, toutes les deux portent des informations significatives du signal. Nous utilisons ensuite la fonction « softmax » pour mettre en évidence le paramètre le plus présent pour chaque fenêtre temporelle selon l'encodage d'Enc. Nous calculons finalement notre score entropique de parole, similairement à l'IS.

2.5 Score entropique de parole

Soit $\boldsymbol{y}$ le vecteur de probabilité correspondant à l'espérance des probabilités d'$Enc(\boldsymbol{x}')$ et $\boldsymbol{y}_t$ le vecteur de probabilité pour un pas de temps ($Enc(\boldsymbol{x}'_t)$). Notez qu'$Enc(\boldsymbol{x}'_t)$ et $\boldsymbol{y}_t$ sont de même dimension. Pour calculer notre score de production ps, nous procédons ainsi :

$$ps(\boldsymbol{x}') = \exp(\rho_{\boldsymbol{x}'_t}(\mathbf{KL}(p(\boldsymbol{y}_t|\boldsymbol{x}'_t)||p(\boldsymbol{y})))) \tag{3}$$

avec $\rho_{\boldsymbol{x}'_t}$ étant une fonction de statistique descriptive et **KL** étant la divergence de Kullback-Leibler.

Pour nos expériences, nous nous sommes limités aux fonctions $\rho_{\boldsymbol{x}'_t}$ suivantes : espérance $\mathbb{E}$, écart-type σ et médiane. Néanmoins, d'autres fonctions statistiques sont possibles. Notez que si $\rho_{\boldsymbol{x}'_t} = \mathbb{E}_{\boldsymbol{x}'_t}$, nous avons la formulation originale de l'IS (Salimans *et al.*, 2016).

Nous avons choisi cette formulation, car elle implique que les productions d'un texte doivent être en même temps : localement stables (idéalement proches d'unités phonétiques) et globalement variées (les participants doivent être capables de prononcer différentes unités sonores). Ainsi, la représentation produite par l'encodeur doit être clairsemée (représentation « sparse »).

3 Expériences

Nous avons essayé plusieurs encodages du signal de la parole dont : les MFCC, les Mel spectrogrammes et la sortie de l'encodeur PASE+. Pour PASE+, nous avons choisi d'utiliser le modèle appris par les auteurs sur 50h d'anglais du corpus Librispeech (Panayotov *et al.*, 2015). Ainsi lors de l'utilisation de ce modèle, nous supposons que l'indice de sévérité est un concept universel (il devrait donc y avoir des similitudes considérables entre l'anglais et le français) et que le modèle encode une représentation suffisamment « bas-niveau » (proche du signal et non de la transcription de la parole) pour être utile aux données françaises.

3.1 Corpus cancer utilisé

Le corpus médical de cancer C2SI (Woisard *et al.*, 2020) contient différentes tâches vocales (comme la lecture ou la description d'une image) pour plusieurs participants (patients et témoins). Toutes les tâches vocales sont en français. Dans nos expériences, nous nous sommes concentrés sur la tâche de lecture afin de disposer du même type d'informations linguistiques à traiter pour chaque locuteur. Un jury de cinq experts cliniques a évalué manuellement la tâche de lecture de chaque participant afin d'obtenir un indice de sévérité. Cet indice représente l'altération pathologique de la production de la parole (Woisard *et al.*, 2022). Pour chaque enregistrement, nous utilisons la moyenne des cinq scores des experts comme score de référence.

Ce corpus contient 82 patients et 25 contrôles et consiste en environ une heure d'enregistrement audio. Nous obtenons ainsi 89 fichiers correspondant aux patients (certains ayant suivi plusieurs sessions espacées de plusieurs jours) et 25 fichiers pour les contrôles. Au total, 114 fichiers ont été enregistrés.

Une évaluation d'une mesure automatique de la parole à l'aide du corpus C2SI a donné une corrélation de Spearman de 0,817 pour la sévérité (Balaguer *et al.*, 2019). Les auteurs ont principalement basé leur score sur la vraisemblance moyenne du phonème attendu, après alignement phonétique.

3.2 Résultats

Un résumé des meilleurs scores obtenus par encodeur est disponible en table 3.2. Le score utilisant le modèle PASE+ donne de meilleurs résultats que ce soit sur les patients uniquement ou sur tous les participants. Il est à noter qu'ici les Mel spectrogrammes nous permettent d'obtenir de meilleurs résultats que les MFCC. Notre méthode est plus performante que celle de (Balaguer *et al.*, 2019) et ne nécessite pas l'utilisation d'un alignement forcé. Notre performance peut être due au débruitage de l'encodeur PASE+, mais également au fait que celui-ci encode plus d'informations de parole que les deux autres paramètres acoustiques utilisés dans nos expériences. Cela est d'autant plus remarquable, que le modèle PASE+ a été appris sur de l'anglais. Ainsi, ce genre de modèles représente des informations de suffisamment bas niveau et génériques pour être appliqué sur d'autres langues.

	MFCC	Mel spectrogrammes	PASE+
Corrélation sur les participants	-0,697	0,800	**-0,868**
Corrélation uniquement sur les patients	-0,637	0,724	**-0,829**

TABLE 1 – Résumé des résultats obtenus pour chaque encodeur.

Un nuage de points pour le résultat utilisant l'encodeur PASE+ est disponible dans la figure 3 : chaque point représente un fichier où un participant lit les mêmes phrases que les autres participants. Nous remarquons que le nuage de points est plus diffus pour les patients avec une faible sévérité [1].

D'une part, bien qu'il soit nécessaire que les patients prononcent (lisent) les mêmes phrases, notre approche ne nécessite pas l'utilisation d'un alignement forcé. D'autre part, le choix de PASE+ (appris avec des techniques de débruitage) nous laisse envisager une certaine robustesse à l'environnement clinique. Ainsi, nous pouvons imaginer l'utilisation de notre approche *in situ*.

4 Conclusion

Dans cet article, nous avons proposé une nouvelle méthode pour évaluer une mesure de la parole semblable à l'indice de sévérité des troubles de la parole chez les patients. Dans notre méthode, nous proposons une adaptation de la sortie d'encodeurs (représentations du signal) pour réaliser un score entropique afin de caractériser la dégradation de la parole des patients. Dans notre meilleure solution, nous utilisons un encodeur (PASE+) pour construire une métrique qui évalue la qualité de la production vocale. Bien que PASE+ ait été appris sur de l'anglais, nos expériences montrent que

1. Notez que vous trouverez sur `https://github.com/vroger11/SAMI` des résultats plus détaillés des expériences, y compris les nuages de points interactifs pour chaque ρ et f_{norm} utilisés avec PASE+ ainsi que le code utilisé.

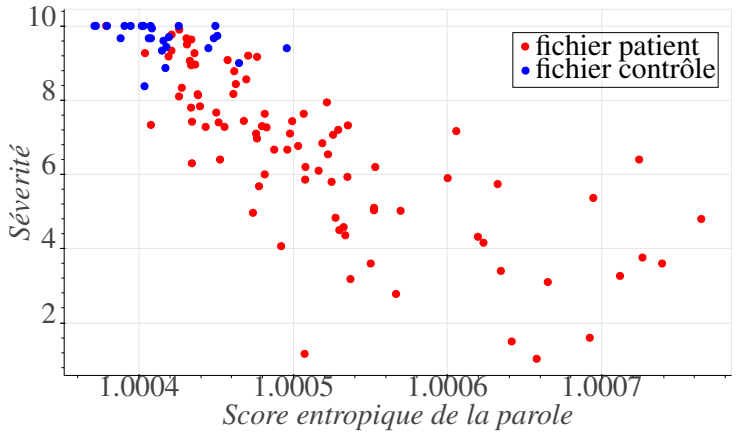

FIGURE 3 – Nuage de points du meilleur score entropique de la parole s'appuyant sur PASE+. Ici f_{norm} utilise le zscore et $\rho_{\boldsymbol{x}'_t}$ utilise la moyenne.

ce modèle est très « général », puisque son utilisation sur une autre langue permet d'obtenir une représentation acoustique plus complète que l'usage direct des MFCC ou des MEL spectrogrammes.

Notre prédiction d'intelligibilité est excellente (corrélation de Spearman de 0,87 avec les experts cliniques), ce qui encourage l'utilisation d'une telle méthode pour les applications médicales au sein de l'hôpital. Ainsi, des réseaux de neurones profonds peuvent être utilisés directement (sans modification) pour des tâches ne disposant pas de suffisamment de données pour réaliser un apprentissage (ou même une mise à jour) des paramètres du modèle. Nos résultats pourraient aussi indiquer que la mesure de la sévérité des troubles de la parole peut être semblable entre certaines langues (ici français et anglais).

Il serait intéressant d'entraîner un modèle français et de voir si nous pouvons obtenir de meilleurs résultats. Nous avons également l'intention d'analyser les représentations intermédiaires produites par plusieurs encodeurs pour voir si nous pouvons récupérer des informations plus fines concernant les mauvaises prononciations.

Références

BALAGUER M., FARINAS J., PINQUIER J. & WOISARD V. (2019). Construction of the automatic Carcinologic Speech Severity Index (C2SI) score. In *31st World Congress of the International Association of Logopedics and Phoniatrics (IALP)*, p. 1–15 : IALP : International Association of Logopedics and Phoniatrics.

ELLINGTON T. D. (2020). Trends in incidence of cancers of the oral cavity and pharynx—united states 2007–2016. *MMWR. Morbidity and Mortality Weekly Report*, **69**.

ENDERBY P. (2013). Disorders of communication : Dysarthria. In *Handbook of Clinical Neurology*, volume 110, p. 273–281. Elsevier.

FANG C., LI H., MA L. & ZHANG M. (2017). Intelligibility Evaluation of Pathological Speech through Multigranularity Feature Extraction and Optimization. *Computational and Mathematical Methods in Medicine*, **2017**, 1–8.

FLETCHER A. R., WISLER A. A., MCAULIFFE M. J., LANSFORD K. L. & LISS J. M. (2017). Predicting intelligibility gains in dysarthria through automated speech feature analysis. *Journal of Speech, Language, and Hearing Research*, **60**(11), 3058–3068.

JANBAKHSHI P., KODRASI I. & BOURLARD H. (2019). Spectral subspace analysis for automatic assessment of pathological speech intelligibility. In *Proc. Interspeech 2019*, p. 3038–3042.

LAARIDH I., KHEDER W. B., FREDOUILLE C. & MEUNIER C. (2017). Automatic Prediction of Speech Evaluation Metrics for Dysarthric Speech. In *Interspeech*, p. 1834–1838, Stockholm, Sweden.

LANDA S., PENNINGTON L., MILLER N., ROBSON S., THOMPSON V. & STEEN N. (2014). Association between objective measurement of the speech intelligibility of young people with dysarthria and listener ratings of ease of understanding. *International journal of speech-language pathology*, **16**(4), 408–416.

MOATTAR M. H. & HOMAYOUNPOUR M. M. (2009). A simple but efficient real-time Voice Activity Detection algorithm. In *17th European Signal Processing Conference*, p. 2549–2553.

PANAYOTOV V., CHEN G., POVEY D. & KHUDANPUR S. (2015). Librispeech : An ASR corpus based on public domain audio books. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 5206–5210, South Brisbane, Queensland, Australia.

POMMÉE T., BALAGUER M., MAUCLAIR J., PINQUIER J. & WOISARD V. (2021). Assessment of adult speech disorders : current situation and needs in French-speaking clinical practice. *Logopedics Phoniatrics Vocology*, p. 1–15.

RAVANELLI M., ZHONG J., PASCUAL S., SWIETOJANSKI P., MONTEIRO J., TRMAL J. & BENGIO Y. (2020). Multi-task self-supervised learning for Robust Speech Recognition. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 6989–6993.

SALIMANS T., GOODFELLOW I., ZAREMBA W., CHEUNG V., RADFORD A. & CHEN X. (2016). Improved Techniques for Training GANs. In D. D. LEE, M. SUGIYAMA, U. V. LUXBURG, I. GUYON & R. GARNETT, Eds., *Advances in Neural Information Processing Systems 29*, p. 2234–2242. Curran Associates, Inc.

SZEGEDY C., VANHOUCKE V., IOFFE S., SHLENS J. & WOJNA Z. (2016). Rethinking the Inception Architecture for Computer Vision. *IEEE - Conference on Computer Vision and Pattern Recognition (CVPR)*.

WALSHE M. & MILLER N. (2011). Living with acquired dysarthria : The speaker's perspective. *Disability and Rehabilitation*, **33**(3), 195–203.

WOISARD V., ASTESANO 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*.

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*, **44**(1), 71–88.

WOISARD V. & LEPAGE B. (2010). Perception of speech disorders : Difference between the degree of intelligibility and the degree of severity. *Audiological Medicine*, **8**, 171–178.