Université Paul Sabatier - Faculté de médecine Toulouse Rangueil, CFUO de Toulouse

Mémoire présenté pour l'obtention du

Certificat de Capacité d'Orthophoniste

par Clémentine JAUNEAU

Evaluation par analyses automatique et perceptive des altérations de la voix de patients atteints de cancers des voies aérodigestives supérieures : vers une réduction de la tâche de lecture ?

Maîtres de mémoire

Virginie WOISARD

Imed LAARIDH

Date de soutenance

Septembre 2019

«La perception de la réalité sonore n’est pas un enregistrement direct de la réalité. C’est une construction mentale opérée à la suite d’un traitement de l’information disponible, contrainte par nos sens ainsi que nos habitudes sélectives » (Gaillard et al., 2007)

REMERCIEMENTS

Je tiens tout d'abord à adresser un grand merci à mes maîtres de mémoire, Virginie Woisard et Imed Laaridh, pour m'avoir permis d'effectuer ce mémoire. Merci à Imed Laaridh pour tout le travail fourni sur les enregistrements. Merci à Virginie Woisard de m'avoir consacré autant de temps, en particulier dans les derniers moments, pour sa bienveillance, ses conseils et ses encouragements.

Merci à ma famille, mes frères, mes parents pour leur soutien et pour m'avoir permis de réaliser ces études. En particulier, un grand merci à mon père pour m'avoir supportée lors de cette dernière année et de m'avoir soutenue jusqu'au bout.

Merci à ma chère promotion 2014/2019 pour cette solidarité tout au long de ces cinq années, pour tous ces bons moments de partage, d'entraide, de rigolade et de dégustation !

Un merci tout particulier à Margaux, Estelle et Mélodie avec qui j'ai tant partagé pendant cinq ans !

Enfin, Charlotte, merci pour ta présence et ton soutien sans faille depuis toutes ces années, dans les moments de joie comme dans les moments plus difficiles.

Et un merci à toutes les personnes qui ont pu, de près comme de loin, participer à ce projet.

TABLE DES MATIERES

1. INTRODUCTION

1.1. Les cancers des voies aérodigestives supérieures

Les cancers des voies aérodigestives supérieures (VADS) ont un taux d'incidence élevé. En 2018, le nombre de nouveaux cas estimé en France est de 13 692 personnes[1] (Institut National du Cancer, 2019). Une évolution favorable est observée quant à l'incidence des cancers des VADS chez l'homme. Au contraire, ce taux augmente chez la femme. Ces changements sont à mettre en corrélation avec la modification des comportements de santé quant à l'exposition aux facteurs de risques tels que la consommation d'alcool et de tabac. Néanmoins, grâce aux progrès dans le domaine de la recherche médicale, la mortalité liée à ce type de cancers recule et un allongement de la durée de vie est notable. De ce fait, un plus grand nombre de personnes vivent avec les séquelles liées aux traitements de ces pathologies ou développent des effets secondaires tardifs. En effet, ces traitements peuvent être particulièrement invalidants engendrant des déficits tels que des troubles de la déglutition, des modifications de la voix et des difficultés dans l'articulation. Les compétences de communication du patient se voient donc altérées par le cancer et/ou le traitement, occasionnant des troubles sur la vie sociale. Par conséquent, l'évaluation de la parole est primordiale, l'évaluation de l'intelligibilité en étant la mesure clé (Ghio et al., 2016).

1.2. Mesure de l'intelligibilité et ses limites

1.2.1. L'intelligibilité

L'intelligibilité est définie comme « le caractère de ce qui peut être facilement compris dans le sens à la fois de la forme et du contenu » (Dictionnaire d'orthophonie, 2018). Autrement dit, l'intelligibilité de la parole est le degré de précision avec lequel un message est compris par un individu, message que l'on peut comprendre en s'appuyant sur des liens logiques et explicites. De ce fait, les patients atteints de troubles de production de la parole ont souvent une plainte sur la perte d'intelligibilité. Perte qui réduit la capacité d'interaction avec les autres et qui participe alors à la diminution de la qualité de vie au niveau communicationnel.

[1] En Annexe 1, le taux d'incidence et de mortalité par régions anatomiques dans les voies aérodigestives supérieures en 2018

1.2.2. L'évaluation de l'intelligibilité

La méthode la plus utilisée par les cliniciens et les orthophonistes pour évaluer l'intelligibilité, encore aujourd'hui, se trouve être l'évaluation perceptive. Elle consiste en la passation de plusieurs tâches, plus ou moins écologiques, pour permettre une mesure de la sévérité de l'atteinte et ainsi permettre d'apprécier le handicap communicationnel du patient, d'un point de vue du degré de compréhensibilité (compréhension du sens du message) ou du degré d'intelligibilité (identification des sons ou des mots produits).

Habituellement, l'évaluation perceptive se présente de la manière suivante : le patient lit une liste de mots et le clinicien transcrit ce qu'il comprend. Les transcriptions sont ensuite comparées à la liste lue et cotées de façon binaire (correct ou incorrect)[2]. Le pourcentage d'éléments correctement reconnus correspond alors au score d'intelligibilité du patient. Néanmoins, cette évaluation perceptive présente des limites qui peuvent ensuite engendrer un biais dans l'élaboration du projet thérapeutique du patient.

1.2.3. Les limites de l'évaluation perceptive

La critique la plus souvent évoquée à l'évaluation perceptive est son caractère subjectif. En effet, dans un contexte clinique, l'examinateur est généralement unique, par conséquent, l'évaluation en est très dépendante. Cela est dû au fait que chaque auditeur possède une représentation de la normalité qui lui est propre et qui dépend de son expérience, âge, langue ainsi que de certains facteurs socio-culturels. C'est ce qui est nommé le « référent interne » de chaque auditeur (Fex, 1992). Dès lors, ce caractère subjectif rend cette évaluation non reproductible.

De plus, la perception de la parole est un mécanisme intégrant à la fois des informations dites de bas niveau (informations provenant du signal, décodage acoustico-phonétique) et des informations dites de haut niveau (contexte de la situation communicationnelle, connaissances des communicants, …). Lorsque nous sommes confrontés à un discours dégradé, ces mécanismes top-down vont nous aider à restaurer le message et optimiser l'intelligibilité, expliquant de ce fait la variabilité intra et inter auditeur (Warren et al., 1970). Ces processus top-down induisent, en outre, un effet de lexicalité c'est-à-dire qu'une séquence sonore

[2] Fontan, 2012

entendue va faire référence à un mot du vocabulaire de l'auditeur[3]. Dans un contexte d'évaluation, un effet d'apprentissage par l'habituation de l'auditeur va se créer. Effectivement, lorsque les mêmes stimuli sont présentés plusieurs fois, l'auditeur finit par en reconnaître et, ainsi, les identifie plus facilement. Tout ceci peut amener à un score d'intelligibilité surévalué.

1.2.4. Les approches par traitement automatique

Avec l'essor des outils technologiques, des approches par traitement automatique de la parole permettraient de fournir des outils objectifs pour l'évaluation de l'intelligibilité. Des méthodes récentes consistent à enregistrer la parole du patient puis à soumettre le signal à un traitement automatique pour obtenir un degré d'intelligibilité des mots ou des phrases produits.

C'est dans ce cadre que s'est inscrit le projet C2SI (Carcinologic Speech Severity Index) qui cherchait à établir de manière automatique un indice de sévérité pour mesurer l'impact d'un trouble de la parole chez les patients traités pour un cancer de la cavité buccale ou du pharynx.

1.3. Projet C2SI

Le projet C2SI (Astésano et al., 2018), a impliqué plusieurs équipes de recherche françaises : le Centre Hospitalier Universitaire de Toulouse, le Laboratoire Parole et Langage d'Aix-en-Provence, le Plateau d'Etudes Techniques et de Recherche en Audition de l'Université Toulouse Jean Jaurès, le Laboratoire Octogone-Lordat de Toulouse, le Laboratoire Informatique d'Avignon de l'Université d'Avignon et des Pays de Vaucluse, et l'Institut de Recherche en Informatique de Toulouse. Il avait pour objectif la création d'un indice automatique de sévérité de la parole applicable sur les pathologies cancéreuses notamment les cancers de la cavité buccale et du pharynx.

Chez les patients traités pour ces cancers, la mesure d'intelligibilité de la parole est peu fiable. Cela est dû à la faible reproductibilité inter-juge des niveaux d'intelligibilité attribués lors des auditions. Par ailleurs, il convient de noter un effet de familiarisation du clinicien à ce type de parole (Warren et al., 1970) ainsi qu'aux tâches proposées au travers des auditions, du

[3] « […] *nous pouvons prédire qu'en français une séquence prononcée [tisk] sera perçue /disk/ en référence au mot « disque »* [...]» (Ghio, 2016)

fait de leur connaissance approfondie et de leur usage répété. Cette double accoutumance a pour conséquence une sous-estimation du niveau de sévérité perçue. De plus, il est à noter un manque d'outils validés pour évaluer l'impact des traitements des cancers de la cavité orale et du pharynx, notamment en ce qui concerne les troubles de la parole.

De ce constat, les équipes impliquées ont émis l'hypothèse que le traitement automatique du signal pourrait être utilisé pour mesurer l'impact d'un trouble de la parole sur les capacités de communication en donnant un indice de sévérité chez les patients traités pour un cancer de la cavité buccale ou du pharynx : le Carcinologic Speech Severity Index (C2SI).

Dans un premier temps, les équipes de recherche ont constitué un corpus de parole composé de nombreuses tâches linguistiques (tenues de voyelles, lecture de texte, production de pseudo-mots, …) recueilli auprès de 129 sujets (87 patients traités pour un cancer de la cavité buccale ou de l'oropharynx et 42 sujets contrôles). Les participants ont donc dû réaliser huit épreuves au total : une épreuve d'intelligibilité, une épreuve de compréhensibilité, trois épreuves de production prosodique et trois tâches de sévérité incluant la lecture d'un extrait du texte de « La chèvre de Monsieur Seguin » d'Alphonse Daudet. Les productions ont ensuite fait l'objet d'une évaluation perceptive et d'un traitement automatique. Les résultats obtenus dans chaque modalité ont été confrontés par la suite, l'objectif final étant d'élaborer un score automatique global.

Une évaluation perceptive constitue une étape à la mise au point d'un index C2SI, index automatique évaluant la dégradation de la communication verbale post-cancer. Dans le cadre de cette évaluation, il est nécessaire d'envisager la parole de ces patients du point de vue de sa réception par un auditeur. D'une part, le recours à l'évaluation perceptive de la parole permet d'obtenir des points de comparaison (Hermes, 1998) pour estimer la validité perceptive d'un indice automatique. D'autre part, le handicap lié aux séquelles des traitements des cancers des VADS concerne particulièrement la réception de cette parole altérée : les difficultés d'articulation dues à la modification des organes de la parole conduiraient à une perte d'intelligibilité du message et une altération de la compréhensibilité du message (Woisard et al., 2013). Dans ces termes, l'intelligibilité renvoie au traitement des informations de bas niveau, tandis que la compréhensibilité évoque des stratégies de haut niveau (compensations, jugements de plausibilité à partir des connaissances préalables de l'auditeur).

Le score automatique C2SI pour chacun des 129 sujets a intégré des paramètres acoustiques de la fréquence de la voix, des scores de vraisemblance automatique sur des tâches

de production de non-mots et de lecture de texte et d'autres modalités de traitement automatique sur la production de non-mots. La modélisation effectuée a produit un coefficient de corrélation de Spearman avec le score perceptif de sévérité à 0,87 (Balaguer-Navarro, 2018). A partir des résultats du traitement perceptif, il a été bien établi que le traitement chirurgical sur la tumeur altère significativement l'intelligibilité et la sévérité de la parole et que le volume tumoral a également un impact sur les performances d'intelligibilité. Ces résultats ouvrent des perspectives sur l'utilisation en pratique clinique du traitement automatique de la production de la parole en cancérologie des voies aérodigestives supérieures.

Lors de la réalisation du corpus, il a été souligné une limitation récurrente. Cette limitation concerne le nombre d'épreuves à réaliser par chaque patient. Ces patients, atteints d'une pathologie cancéreuse, présentent une fatigabilité pouvant les contraindre à abandonner une tâche de production de parole dans le cadre d'un protocole d'évaluation. Il apparaît donc opportun d'estimer si l'ensemble du protocole d'enregistrement est nécessaire ou pas au bon fonctionnement des outils automatiques d'évaluation.

1.4. Objectifs

Ainsi, nous posons l'hypothèse que l'analyse par traitement automatique d'un échantillon de parole sur de la lecture de phrases, différentes du texte de « La chèvre de Monsieur Seguin » d'Alphonse Daudet, amène à des résultats similaires. De plus, nous faisons l'hypothèse que l'analyse automatique d'un temps de parole réduit, en termes de phrases, pourrait conduire à un résultat fiable.

L'objectif de ce travail vise à estimer dans quelle mesure il est possible de mener à bien une évaluation de la sévérité d'altération de la parole au travers d'un protocole simplifié en termes de tâches pour le patient.

2. MATERIEL ET METHODE

2.1. Population d'étude

La population étudiée comprend des patients atteints de cancers des voies aérodigestives supérieures (cavité buccale ou pharynx). Tous les patients ont subi un traitement dédié consistant en une chirurgie, et/ou une radiothérapie, et/ou une chimiothérapie. Les enregistrements sélectionnés sont rétrospectifs, sélectionnés dans la base de données des enregistrements réalisés dans le parcours de soins de chaque patient (suivi de rééducation). Ils doivent contenir dans tous les cas : la lecture du texte de « La chèvre de Monsieur Seguin », les 10 phrases de Combescure (en lecture ou répétition) et de la parole spontanée.

2.2. Sélection de l'échantillon

Les enregistrements constituant le corpus ont été obtenus à partir de la base de données d'enregistrements des patients dans le cadre de leur suivi de rééducation. Ils ont été réalisés sur les différents sites de l'Institut Universitaire du Cancer de Toulouse (IUCT) : l'Oncopôle et l'Hôpital Larrey de Toulouse. Les enregistrements sont réalisés dans une cabine insonorisée avec microphone à l'Oncopôle, tandis qu'à l'Hôpital Larrey, les enregistrements ont lieu dans une pièce non insonorisée avec casque-micro. Les fichiers audio sont tous au format WAVE.

Une première écoute de 180 enregistrements a permis d'éliminer les doublons et les enregistrements incomplets.

Une seconde écoute a eu pour but de classer, de manière subjective, les enregistrements retenus, en 5 catégories de sévérité : sévérité grave, grave-médium, médium, médium-légère, légère. Ce classement, réalisé par un auditeur unique, est donc sujet aux limites explicitées précédemment. De plus, la préparation des stimuli (présentée en 2.4.2) a amené une habituation à ces stimuli et en particulier aux différents locuteurs, cela ayant pour risque une surévaluation de l'intelligibilité. De ce fait, un premier test d'évaluation perceptive a été réalisé par un panel d'auditeurs, auditeurs naïfs mais avec une certaine sensibilité à la problématique, à savoir 9étudiantes en 1ère année d'orthophonie au Centre de Formation de Toulouse. Cette écoute avait pour but de vérifier si les différents enregistrements retenus couvraient les différents niveaux

de sévérité. La tâche consistait, après écoute du fichier audio, à cocher la case du degré de sévérité de l'intelligibilité perçu[4]. Les résultats sont présentés dans l'annexe 3.

Suite à cette étape préliminaire, les enregistrements finaux retenus constituent un corpus de 51 enregistrements répartis comme suit selon les catégories de sévérité : 10 « grave », 11 « grave-médium », 11 « médium », 9 « médium-légère » et 10 « légère ». Afin de garantir l'anonymat, un code a été attribué à chaque sujet.

2.3. Données cliniques

En parallèle, des données individuelles et cliniques ont été collectées pour l'ensemble des sujets :

- Age lors de l'enregistrement ;
- Sexe ;
- Site du lieu d'enregistrement ;
- Région anatomique touchée par la pathologie cancéreuse ;
- Critères T (« taille ») et N (« extension ganglionnaire ») de la classification TNM de la tumeur[5] ;
- Type de traitement

2.4. Stimuli

2.4.1. Stimuli sélectionnés

Les épreuves qui ont été conservées pour chaque patient sont l'épreuve de lecture d'un texte, et celle de lecture de phrases.

La tâche de lecture consiste en la lecture du premier paragraphe de « La chèvre de Monsieur Seguin » d'Alphonse Daudet :

« Monsieur Seguin n'avait jamais eu de bonheur avec ses chèvres, il les perdait toutes de la même façon, un beau matin elles cassaient leur corde, s'en allaient dans la montagne et là-haut

[4] Feuille consigne en Annexe 2
[5] Tableau en Annexe 4

le loup les mangeait, ni les caresses de leur maître, ni la peur du loup, rien ne les retenait. C'était paraît-il des chèvres indépendantes voulant à tout prix le grand air et la liberté. »

Ce texte n'est pas équilibré phonétiquement (Sicard et al., 2017) néanmoins il est couramment utilisé dans la pratique clinique pour l'évaluation des troubles de la voix et de la parole. Selon le lieu de l'enregistrement, les patients n'ont pas eu la globalité du paragraphe à produire, une partie ayant été tronquée (« […] et là-haut le loup les mangeait […] C'était paraît-il des chèvres … et la liberté. »)

La tâche de lecture de phrases consiste en la lecture de 10 phrases provenant des listes de phrases de Combescure.

En 1981, P. COMBESCURE a publié dans la Revue d'Acoustique, vingt listes composées chacune de dix phrases de longueur variable. Ces listes sont dites phonétiquement équilibrées, c'est-à-dire que les mots sont choisis de façon à ce que le taux de récurrence de chaque phonème dans la phrase soit le plus proche de celui présent dans la langue française.

De nombreux auteurs (Schuster et Stelzle, 2012 ; Kraaijenga et al., 2016) recommandent l'utilisation d'un matériel équilibré qui inclut tous les phonèmes de la langue testée. Les phrases de Combescure répondent donc à cette exigence. Par ailleurs, l'utilisation de phrases s'approche le plus possible des situations conversationnelles qu'un individu peut rencontrer dans la vie quotidienne d'où le choix de garder ce support.

Dans tous les enregistrements, les phrases utilisées sont issues de la première liste :

- Il se garantira du froid avec un capuchon.
- Annie s'ennuie loin de ses parents.
- Les deux camions se sont heurtés de face.
- Un loup s'est jeté immédiatement sur la petite chèvre.
- Dès que le tambour bat, les gens accourent.
- Mon père m'a donné l'autorisation.
- Vous poussez des cris de colère ?
- Ce petit canard apprend à nager.
- La voiture s'est arrêtée au feu rouge.
- La vaisselle propre est mise dans l'évier.

2.4.2. Préparation des stimuli

Pour permettre le traitement et l'analyse des épreuves sélectionnées, il a été nécessaire de procéder à une segmentation de chaque fichier audio afin de ne conserver que les segments voulus. De plus, la segmentation a permis de supprimer les phénomènes parasites (consignes, bruits annexes) pour isoler au maximum la voix du patient. Nous avons également pris le parti de supprimer la fin du texte de « La chèvre Monsieur Seguin » du fait de son absence pour certains locuteurs.

Cette segmentation s'est également accompagnée d'une transcription du message produit par le patient afin de permettre le traitement automatique. Ces deux actions ont été réalisées sur le logiciel PRAAT.

Enfin, tous les fichiers audio ont été amplifiés afin de limiter un biais lié à une intensité faible de l'enregistrement.

2.5. Traitement perceptif

2.5.1. Auditeurs

Pour pallier les limites d'une écoute perceptive et diluer l'effet du « référent interne » de chaque auditeur, nous avons décidé de monter un jury d'écoute.

Nous avons fait le choix de constituer ce jury d'écoute avec des auditeurs naïfs, c'est-à-dire non familiarisés avec l'écoute de la parole altérée. L'objectif était d'éviter au mieux les effets de processus de haut niveau et les effets d'habituation à la parole pathologique qui sont habituellement rencontrés chez les auditeurs experts et qui peuvent amener à une surévaluation de l'intelligibilité. De plus, ces auditeurs étaient natifs de langue française avec un bon niveau d'audition (afin de s'assurer d'une bonne perception sensorielle). Notre jury d'écoute était donc composé de 30 auditeurs (19 – 45 ans, 16 femmes et 14 hommes).

2.5.2. La tâche de perception

Chaque phrase a été randomisée de sorte que chacune d'entre elles soit évaluée par trois auditeurs différents. De plus, elles étaient présentées dans un ordre aléatoire. De même, pour le texte de « La chèvre de Monsieur Seguin ».

Chaque auditeur a dû évaluer un total de 51 phrases et 5 textes de « La chèvre de Monsieur Seguin » (trois auditeurs ont eu 1 texte en plus pour que chaque texte soit évalué par trois auditeurs différents).

Les auditeurs ont dû noter sur une échelle allant de « aucune altération » (correspondant à 0) à « altération sévère » (correspondant à 10), l'altération globale du signal sonore[6]. Pour cela, l'auditeur devait positionner sur une ligne horizontale non graduée une marque entre 0 et 10. Le score obtenu par locuteur équivaut à la moyenne de chacune des évaluations perceptives. Nous avons dissocié le score obtenu pour la lecture des phrases de Combescure à celui de la lecture de texte afin de pouvoir les comparer.

2.6. Traitement automatique

En ce qui concerne l'analyse par traitement automatique, chaque enregistrement a été soumis à un traitement basé sur un alignement automatique contraint par le texte. L'alignement automatique contraint consiste, à partir d'un signal de parole et de sa transcription en mots, à détecter les frontières de chaque phonème présent dans la phrase ou le texte prononcé. Le système prend en entrée le signal de la parole (sous forme de paramétrisation), la transcription de la parole produite, un lexique phonetisé contenant les différents mots du texte ainsi qu'une modélisation des phonèmes du Français appris automatiquement sur un grand corpus d'enregistrement radiophonique.

Cet alignement automatique permet aussi de calculer un score de vraisemblance entre chaque segment (phonème) produit par le locuteur et le modèle générique du phonème donné au système automatique en entrée. Ce score permet de caractériser si la production du locuteur a été proche ou pas du modèle générique. Plus ce score augmente, plus la production du locuteur est éloignée du modèle de la parole et donc de la normale.

Les scores extraits seront utilisés afin d'avoir des scores par locuteur et par tâche de production (phrases et texte de Monsieur Seguin).

[6] Fiche d'évaluation avec consignes en Annexe 5

3. RESULTATS

3.1. Description de la population

Notre échantillon se compose de 51 patients. La répartition femmes/hommes est de 20 femmes (39%) et de 31 hommes (61%). L'incidence des cancers des voies aérodigestives est supérieure chez l'homme que chez la femme (10 055 cas chez l'homme et 3 637 cas chez la femme estimés en 2018[7]). Notre population d'étude est donc représentative de la population au vue de cette incidence. L'âge des patients a été calculé en fonction de la date d'enregistrement de leur suivi, les enregistrements récupérés s'étalant de 2006 à 2018. La moyenne d'âge des patients est de 61 ans (de 31 à 87 ans).

La répartition selon le lieu de l'enregistrement est de 19 personnes provenant de l'Hôpital Larrey de Toulouse (37%) et 32 personnes de l'Oncopôle (63%).

Sur les 51 patients, 27 avaient une atteinte de l'oropharynx (53%), 22 avaient une atteinte de la cavité buccale (43%) et 2 avaient une atteinte du rhinopharynx (4%). Le tableau 1 donne en détail les localisations des tumeurs des patients. L'annexe 6 fournit la répartition anatomique des tumeurs en fonction du sexe.

Tableau 1 : détail des localisations des tumeurs des patients

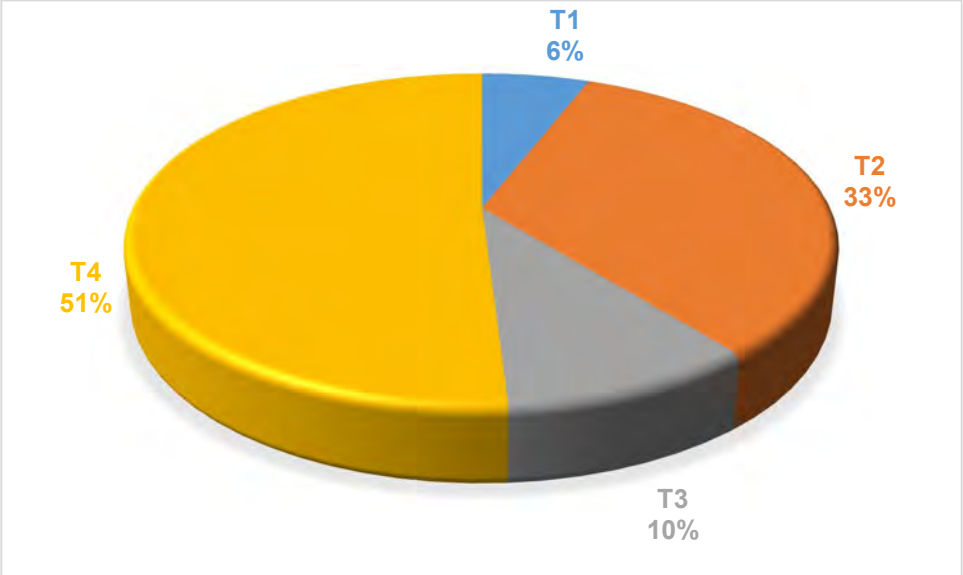
	Cavité buccale	Oropharynx	Rhinopharynx
Langue	10		
Plancher buccal	6		
Trigone rétromolaire	2		
Gencive / Mandibule	2		
Cavité buccale étendue	2		
Amygdale		13	
Base de langue		9	
Oropharynx étendu		3	
Voile du palais		2	
Cavum			2
Total	22	27	2

3 patients présentaient une tumeur classée T1, selon la classification TNM. La tumeur de 17 patients était classée en T2. 5 patients présentaient une tumeur T3 et enfin, le groupe de

[7] Institut National du Cancer (2019). Synthèse – Estimations nationales de l'incidence et de la mortalité par cancer en France métropolitaine entre 1990 et 2018

patients le plus important était celui des tumeurs classées en T4 avec 26 patients. La figure suivante présente cette répartition.

Figure 1 : répartition en fonction de la taille de la tumeur (critère « T » de la classification TNM)



Au sujet des différents traitements entrepris, la chirurgie est généralement le traitement de première intention d'un cancer des VADS (La Ligue contre le Cancer, & Institut National du Cancer, 2018). La majorité des patients de notre étude a subi une chirurgie soit 42 patients sur 51 (82%). Cette chirurgie a pu être proposée de manière isolée ou associée à de la radiothérapie et/ou chimiothérapie. Les 9 patients qui n'ont pas eu de chirurgie ont bénéficié d'une radiothérapie accompagnée ou non de chimiothérapie (18%). L'annexe 7 présente le détail des traitements entrepris en fonction de la région anatomique de la tumeur ainsi qu'en fonction de la taille de la tumeur.

3.2. Traitement perceptif

En préambule à toute analyse des résultats, nous avons vérifié, dans le cas de notes qui semblaient discordantes, que l'auditeur en cause n'avait pas réalisé une notation inverse ou que l'entrée des notes n'avait pas été décalé.

Pour rappel, chaque phrase a été notée trois fois par des auditeurs différents. De plus, chaque enregistrement d'un locuteur contenait 10 phrases. Soit au total, 30 notes pour un locuteur sur les phrases de Combescure. Auxquelles il faut rajouter les trois notes de l'extrait

de texte. Si nous prenons en compte le nombre de locuteurs pour chaque groupe de sévérité, le nombre total de notes par catégorie est compris entre 270 et 330 notes pour les phrases de Combescure, et entre 27 et 33 notes pour l'extrait de texte.

3.2.1. Moyenne des scores de sévérité

Les moyennes des scores obtenues par phrase, en fonction des groupes de sévérité, montrent des valeurs décroissantes du groupe de sévérité « grave » au groupe de sévérité « légère ». Aucune des phrases ne dérogent à cette décroissance, indiquant que globalement les différentes phrases de Combescure conduisent à des résultats équivalents en termes d'affectation aux groupes de sévérité.

A noter, les valeurs élevées des écart-types soulignant la variabilité des notes attribuées par les auditeurs. Une analyse de cette variabilité est réalisée dans le point suivant (3.2.2.).

Cette même décroissance des valeurs est observée si le calcul de la moyenne prend en compte les notes pour l'ensemble des phrases de Combescure. Elle se vérifie également si nous considérons l'extrait de texte de « La chèvre de Monsieur Seguin » (valeurs indiquées dans les deux dernières lignes du tableau). L'ensemble des phrases de Combescure mènent à des résultats similaires à ceux du texte.

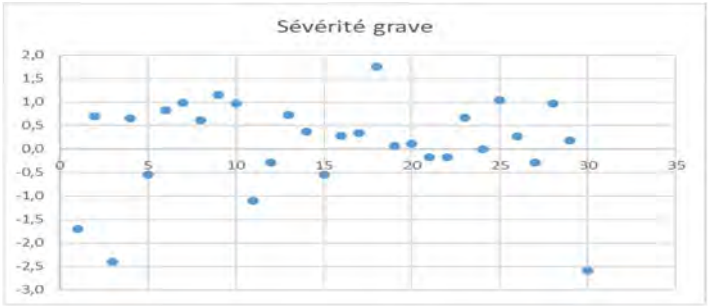
Tableau 2 : Moyenne et écart-type (entre parenthèses) des notes pour chaque phrase, les phrases en totalité et le texte de « La chèvre de Monsieur Seguin » par groupe de sévérité :

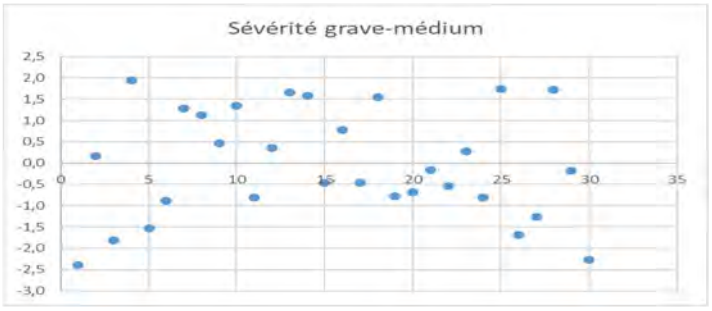
	Sévérité grave	Sévérité grave-médium	Sévérité médium	Sévérité médium-légère	Sévérité légère
Phrase 1	8,80 (1,62)	7,30 (2,54)	7,25 (2,23)	4,85 (2,62)	1,50 (1,60)
Phrase 2	7,35 (2,65)	6,07 (3,31)	5,47 (2,82)	3,64 (2,06)	1,60 (2,12)
Phrase 3	8,85 (1,16)	6,48 (2,63)	5,17 (2,83)	4,31 (2,79)	1,07 (1,47)
Phrase 4	8,33 (1,65)	7,19 (2,15)	6,71 (2,10)	3,79 (2,06)	1,60 (1,80)
Phrase 5	9,03 (1,27)	7,77 (2,16)	6,15 (2,44)	4,89 (2,73)	2,54 (2,46)
Phrase 6	7,51 (2,36)	5,64 (2,89)	5,05 (2,93)	2,46 (2,16)	1,10 (1,58)
Phrase 7	8,88 (1,33)	7,61 (2,36)	7,16 (2,90)	4,83 (3,32)	2,62 (2,35)
Phrase 8	7,27 (2,45)	6,34 (2,64)	6,03 (2,74)	3,81 (2,43)	1,50 (1,40)
Phrase 9	6,88 (2,85)	5,50 (2,72)	3,99 (2,66)	2,84 (2,28)	0,81 (1,15)
Phrase 10	7,43 (2,41)	5,66 (2,36)	5,72 (2,74)	2,44 (2,41)	1,10 (1,43)
Phrases en totalité	8,03 (2,18)	6,56 (2,66)	5,87 (2,79)	3,79 (2,63)	1,54 (1,85)
Texte Mr Seguin	8,41 (1,41)	6,12 (2,34)	5,58 (2,40)	3,38 (2,06)	1,78 (1,87)

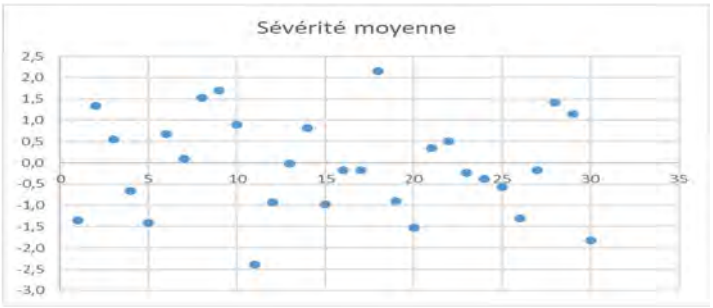
3.2.2. Analyse des différences de perception par auditeur

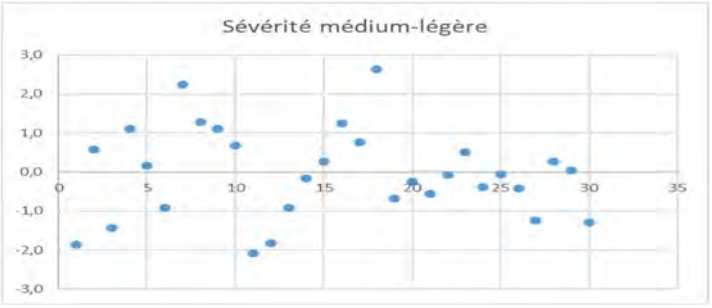
Il est intéressant de déterminer si certains auditeurs surévaluent ou sous-évaluent de façon systématique la mesure de sévérité. Pour cela, nous avons calculé par auditeur, pour chaque groupe de sévérité, la moyenne de leur notation. Cette moyenne est comparée à la moyenne générale du groupe de sévérité considéré. Nous obtenons la figure suivante représentant, par auditeur, les écarts à la moyenne générale, et cela, pour chaque groupe de sévérité. Les valeurs négatives indiquent une sous-évaluation, et les valeurs positives une surévaluation, au regard de la moyenne générale.

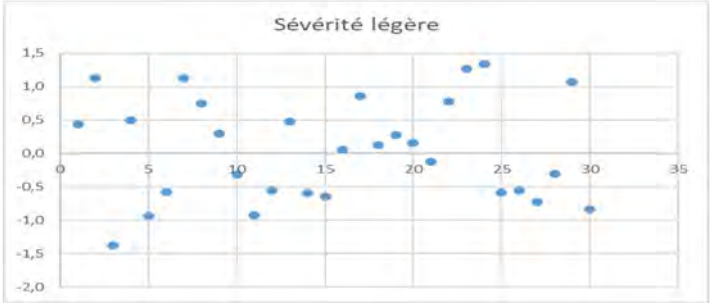
Figure 2 : Différences de perception par auditeur











Nous constatons que seuls quelques auditeurs (plus particulièrement les auditeurs A1, A3 et A30), sous-évaluent de façon systématique le niveau de gravité. Inversement, un auditeur (A18) surévalue systématiquement le niveau de sévérité, au regard de la moyenne générale. Pour ces raisons, l'ensemble des données ont été conservé pour la suite des analyses.

3.2.3. Analyse relative à l'étendue des scores de perception

Nous avons réalisé une analyse de l'étendue des valeurs en fonction des groupes de sévérité. La valeur de l'étendue donne une valeur de la distribution des scores obtenus. Elle correspond à la différence entre la note maximale et la note minimale des scores attribués par chaque auditeur aux stimuli. Pour rappel, chaque phrase et texte de chaque locuteur a reçu trois notes de trois auditeurs différents. Une valeur basse de l'étendue signifie que les auditeurs ont fourni des scores relativement groupés. Inversement, une valeur élevée indique des différences de perception importantes d'un auditeur à l'autre.

Le tableau suivant permet une première visualisation des données. C'est un tableau synthétique regroupant les valeurs des étendues par groupe de sévérité. Le code couleur attribué repose sur une échelle graduée allant de 0 (vert foncé) à 10 (rouge), correspondant à l'échelle des notes. La première colonne correspond au locuteur et les colonnes suivantes correspondent à un stimulus (les 10 phrases de Combescure et le texte de « La chèvre de Monsieur Seguin »).

Au-delà des valeurs indiquées dans le tableau, le codage couleur permet d'avoir une vision globale de la variabilité des résultats en termes de variabilité de la perception et donc des scores attribués à un même couple locuteur-phrase par les différents auditeurs.

Tableau 3 : L'étendue des scores en fonction du degré de sévérité de l'intelligibilité

		Tâches										
		P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Txt
GRAVE	L1	1.5	3	1.8	1.8	1.8	1.8	0.9	3.6	1.4	8.4	1.3
	L2	1.5	1.1	1.4	3.8	2.5	5.4	1.6	4.9	1.5	2.9	0.8
	L3	0.2	6.6	0.1	0.9	0.7	3.7	1.2	7.9	2.8	2.7	4.2
	L4	0.5	1.1	0.7	2.6	0.3	5.1	1.4	4.2	6.1	4	0.1
	L5	0.5	1.8	0.7	3.2	1.2	4.4	0.3	0.3	0.1	0.6	0.2
	L6	2.1	0.2	1.9	3.2	3.4	7.6	4.2	1.7	3.8	5	0.3
	L7	2	9.6	1.7	5	3.8	6.1	2.1	5.8	6	4.3	2.4
	L8	0.3	0.7	0.9	0.4	0.8	1.5	1.9	1.7	0.7	1.2	2
	L9	7.4	2.9	4.5	2.1	0.3	2.7	5.2	1.1	1.6	4.1	2.1
	L10	2.9	7.6	2.5	2	1.7	3.9	0.7	5.2	5	3.8	1.8
GRAVE-MEDIUM	L1	1.4	0.2	1.3	2.4	0.2	2.7	4.8	0.9	5.6	4.9	2.1
	L2	2.6	6.4	2.8	2.4	6.6	2.1	5.7	7.3	4.1	2.7	4.5
	L3	3.9	3.6	3.9	0.7	4.4	8.7	7.7	2	1.9	2.7	4.1
	L4	0.5	3.9	5.4	6.2	4.6	1	2.2	3.5	5.4	1.8	1.3
	L5	4.2	6.8	5.7	2.2	4.1	0.8	1.9	3.5	5.4	3.7	6.6
	L6	8.1	3.9	5.7	4.8	3	6	1.5	3.4	3.5	3	2.5
	L7	4.7	7.4	7.2	2.9	2.4	3.2	0.4	2.2	3.6	6.8	5.1
	L8	4.4	1.7	2.9	4.6	0.1	4.5	0.6	1	1.3	6.8	4.9
	L9	1.8	6.9	1.8	0.8	5.9	7.5	1	2.9	1.9	2.6	4.8
	L10	1.3	9.8	3.4	3.8	1.6	2.3	4.3	1	4	5.4	0.6
	L11	1.4	2.3	4.8	0.7	2.1	3.6	0.4	1.5	4.6	1.4	1.5
MEDIUM	L1	2.4	1.8	4	3.5	2.6	5.3	4.4	1.9	5.2	6	2.3
	L2	5.4	6.1	6.7	1.4	2.9	6.1	3.1	2.2	7.6	2.1	2.7
	L3	2.6	4	8.2	3.9	1.5	2.5	1.8	2.2	8.3	2.2	1.8
	L4	1.9	4.7	2.5	1.8	2.2	5.9	0.4	1.9	4.8	2.9	2.4
	L5	3.5	1.3	5	4.3	4.2	1.3	3.8	3.5	1.9	3.9	7
	L6	2.1	5.2	3.9	3.2	1.9	2.5	0.9	3.1	4.6	4.3	3.2
	L7	4.5	7.2	1.6	1.8	0.9	8	1.7	2.8	6.1	4.8	3.8
	L8	1.6	2.9	0.7	4.5	4.8	5.3	3.2	3.9	1.5	4.6	7.2
	L9	1.1	6.1	1.1	3.7	4.8	4	7.7	1	2.5	3.5	5.5
	L10	7.7	5.8	3.9	2.5	1.8	3.8	4.6	6.7	3.5	4.2	5.1
	L11	0.2	2	2	1.8	4	4.8	1.4	3.1	2.5	3.9	2.7
MEDIUM-LEGER	L1	3.5	1.3	4.1	1.2	1.9	7.9	6.2	1.1	3.1	2.4	4.3
	L2	3.4	5.5	1.8	4	4.2	3.1	1.8	3	1.6	0.9	2.8
	L3	4.4	2.2	1.6	0.9	1.3	5.3	8.4	5.1	3	5.6	2.4
	L4	4.6	3.7	2	2.3	1.5	1.7	8.9	3.1	1.8	4.2	5.5
	L5	3	3.1	2.6	4.9	4	0.2	6.2	0.5	3.7	5.8	5.9
	L6	3.3	1.7	6.7	2.9	2.8	1.3	6.1	3.5	2.3	1.4	3.4
	L7	2.7	4.5	0.4	1.6	0.7	2.2	0.3	2.9	2.9	3.9	5.7
	L8	1.4	5.3	5.1	4.5	1.6	2.6	3	2.8	3.9	7.4	5.1
	L9	7.8	3.1	2.1	1.7	3.4	2.3	4.5	3.2	4.1	3.8	1.5
LEGER	L1	2.9	3	4.8	3.2	1.2	1	2.1	3.7	2.9	5.7	6.3
	L2	0.7	0.4	0.8	0.5	0.3	0.6	0.3	0.1	0.1	0.5	0.6
	L3	2.6	1.4	2.8	3.5	2.2	4.7	1.5	5.1	0.7	2	1.7
	L4	1.9	2.5	2.2	0.5	3.5	2.3	0.8	1.2	1.3	0.3	2
	L5	5.6	2.1	0.1	4.6	1.5	6	3	1.1	5.3	2.2	0.9
	L6	0.3	0.5	0.1	0.1	0	0.1	0.2	0.2	0.2	0.2	1.4
	L7	0.7	3.6	0	0.3	0.4	0.1	3.6	1.2	0.1	0.2	1.7
	L8	2.5	0.8	2.1	2.6	4.2	0.6	7.7	3.1	1	1.4	2.3
	L9	2.6	3.1	1.5	2.1	4.7	0.5	3.3	1	1.6	0.9	1.2
	L10	2.1	1.8	5	5.5	5.4	1.1	2.6	1	1.4	3.4	3.2

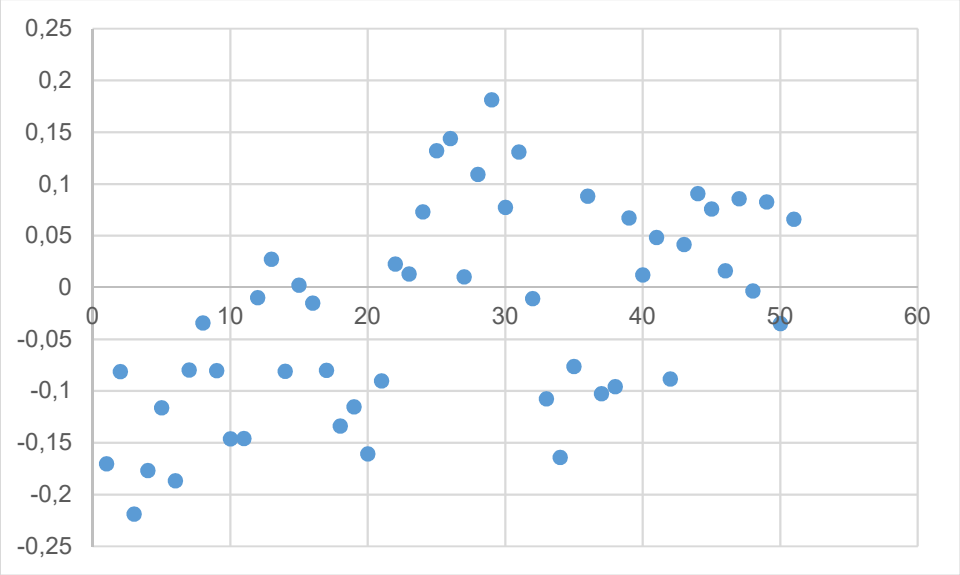
Ce tableau présente un patchwork de couleur qui rend sa lecture et son interprétation difficile. Ce rendu illustre la faible reproductibilité des résultats, au travers de la mesure de l'étendue de l'écoute perceptive, du fait de l'écoute par différents auditeurs et de sa subjectivité. Ce fait constitue une limite à l'évaluation perceptive. Elle est couramment soulignée au travers d'étude (Warren et al., 1970 ; Fontan, 2012) et constitue un des arguments justifiant la nécessité d'une analyse par traitement automatique.

L'examen de ce tableau permet cependant d'extraire deux tendances relatives à la couleur dominante. Une première tendance est constituée des deux groupes de sévérité extrêmes (« grave » et « légère ») qui ont majoritairement une teinte verte dominante. La seconde tendance présente majoritairement une teinte rouge-orange et comprend les trois groupes

médians. Cette tendance se vérifie aussi si nous ne considérons que la colonne correspondant aux valeurs d'étendue pour le texte de « La chèvre de Monsieur Seguin ». Cela traduit donc la grande variabilité des résultats. Ainsi qu'une difficulté plus grande de noter des locuteurs présentant des troubles de production de la parole de sévérité intermédiaire, les groupes de sévérité extrêmes faisant la quasi-unanimité parmi les auditeurs.

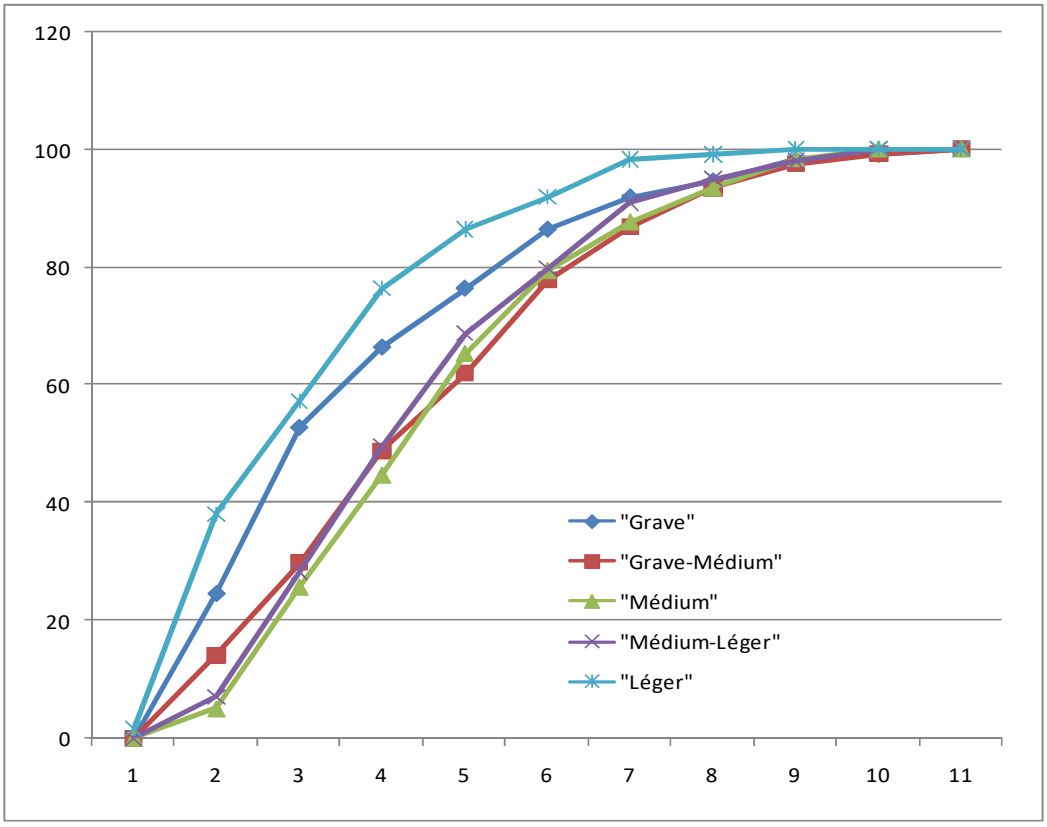
Malgré cette variabilité, nous avons pris l'option de ne pas supprimer les valeurs extrêmes, leur suppression ne modifiant pas significativement les résultats. Pour vérifier ce point, nous avons calculé la moyenne des phrases en excluant les notes extrêmes (moyenne « pondérée »). La figure suivante montre que les différences entre la moyenne globale et la moyenne « pondérée » de l'ensemble des phrases par locuteur sont comprises entre -0,2 et 0,2 soit une différence minime n'influençant pas la tendance générale.

Figure 3 : Différence entre moyenne globale et moyenne « pondérée » des notes des phrases par locuteur



Une autre illustration (Figure 4) de ces résultats consiste à réaliser des courbes de fréquences cumulées par groupe de sévérité en fonction des valeurs d'étendue regroupées en classes. Les 10 classes sont définies avec un pas de 1. La classe 0 est définie pour les valeurs d'étendue nulles, la classe 1 pour des valeurs d'étendue comprises entre 0 (exclu) et 1 (inclus), jusqu'à la classe 10 de valeurs comprises entre 9 (exclu) et 10 (inclus). Les fréquences sont ensuite données en %.

Figure 4 : Courbes de fréquence cumulée par groupe de sévérité en fonction des valeurs d'étendue



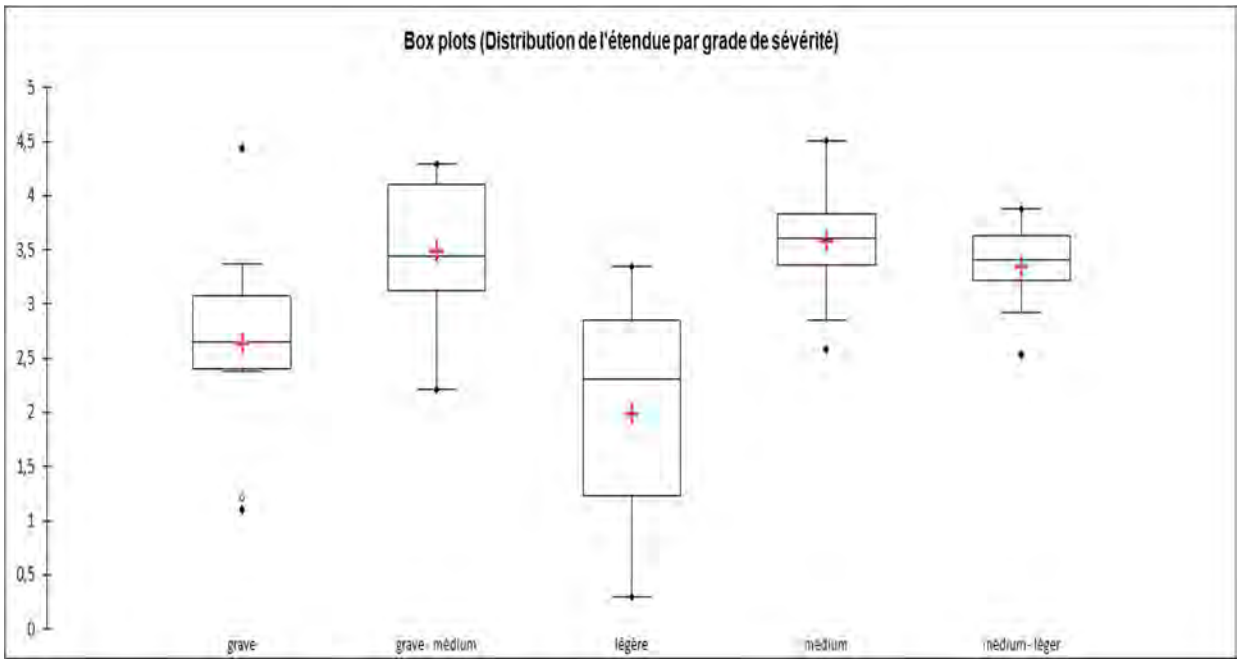
Cette représentation montre clairement deux types de courbe. Les deux courbes bleues correspondent aux groupes extrêmes, les trois autres regroupent les trois groupes de sévérité intermédiaire. Pour les groupes de sévérité « grave » et « légère », plus de 50% des valeurs d'étendue sont inférieures à 2,5 alors que cette valeur est voisine de 4 pour les trois groupes de sévérité intermédiaire.

3.2.4. Valeur moyenne de l'étendue

Nous avons établi la valeur moyenne de l'étendue, tous scores confondus, en fonction des différents groupes de sévérité. Cette analyse permet de minimiser les valeurs extrêmes (qu'elles soient faibles ou élevées) et de mettre en perspective, de façon plus globale, la variabilité des scores.

Une représentation en box plots donne une visualisation de l'ensemble des données au travers des paramètres statistiques de moyenne, de médiane, de quartiles et de valeurs minimale et maximale (Figure 5).

Figure 5 : Distribution de l'étendue par grade de sévérité



Nous pouvons noter que les groupes de sévérité extrême présentent les valeurs moyennes les plus faibles. Cela indique globalement que les auditeurs attribuent des scores groupés aux locuteurs de ces deux groupes de sévérité et ce, toutes tâches confondues. En ce qui concerne le groupe de sévérité « grave », il convient de remarquer que, si l'on excepte les notes extrêmes, les valeurs sont bien groupées autour de la valeur médiane. Paradoxalement, dans le groupe de sévérité « légère », si la valeur moyenne est la plus faible, nous observons néanmoins une plus grande dispersion des valeurs. Cette plus grande variabilité est difficile à interpréter. Elle pourrait être associée à la réticence d'auditeurs naïfs à attribuer de bons scores sachant que ce sont des voix de patients après un cancer des VADS.

En ce qui concerne les groupes « médium », « grave-médium » et « médium-légère », les valeurs moyennes, plus élevées que celles des groupes de sévérité précédents, sont équivalentes les unes aux autres (à savoir 3,58 ; 3,49 et 3,35 respectivement). Cela indique, comme pressenti, une plus grande dispersion des scores attribués selon les auditeurs.

3.2.5. Corrélation entre le texte de « La chèvre de Monsieur Seguin » et les phrases de Combescure

Les résultats précédents indiquent que, à partir de notes issues d'auditeurs naïfs, et au-delà de la variabilité de cette notation, les phrases de Combescure et l'extrait de texte conduisent à des résultats similaires en termes d'évaluation du niveau de sévérité. Il est rappelé ci-dessous le contenu des phrases ainsi qu'indiqué leur nombre de mots et de syllabes.

- P1 : Il se garantira du froid avec un capuchon. (8 mots, 14 syllabes)
- P2 : Annie s'ennuie loin de ses parents. (7 mots, 9 syllabes)
- P3 : Les deux camions se sont heurtés de face. (8 mots, 10 syllabes)
- P4 : Un loup s'est jeté immédiatement sur la petite chèvre. (10 mots, 16 syllabes)
- P5 : Dès que le tambour bat, les gens accourent. (8 mots, 10 syllabes)
- P6 : Mon père m'a donné l'autorisation. (7 mots, 11 syllabes)
- P7 : Vous poussez des cris de colère ? (6 mots, 8 syllabes)
- P8 : Ce petit canard apprend à nager. (6 mots, 10 syllabes)
- P9 : La voiture s'est arrêtée au feu rouge. (8 mots, 11 syllabes)
- P10 : La vaisselle propre est mise dans l'évier. (8 mots, 11 syllabes)

Seule la phrase P7 est une phrase interrogative dont la prosodie est suggérée par la présence du point d'interrogation.

Quant à l'extrait de texte de « La chèvre de Monsieur Seguin », celui-ci est composé de 47 à 53 mots (selon l'extrait proposé au patient) soit 64 à 71 syllabes.

Au regard de notre objectif de réduction des tâches à réaliser par le patient dans le cadre de son évaluation clinique, une confrontation des scores obtenus pour le texte de « La chèvre de Monsieur Seguin » est réalisée avec ceux des phrases de Combescure. Le facteur de corrélation obtenu entre l'ensemble des 10 phrases de Combescure et l'extrait de texte est de 0,85.

L'estimation du facteur de corrélation a également été réalisée par phrase (notée de P1 à P10) ou par combinaison de deux phrases. L'objectif est de déterminer dans quelle mesure l'utilisation d'un nombre réduit de phrases peut conduire à une estimation correcte du niveau de sévérité. Un modèle de régression linéaire a été appliqué et les facteurs de corrélation sont indiqués dans le tableau 3. Une valeur de ce facteur de corrélation égale ou supérieure à 0,8 indique un bon niveau de corrélation.

Tableau 3 : Corrélation entre le texte de « La chèvre de Monsieur Seguin » et les phrases de Combescure :

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Phrase seule	0,66	0,7	0,68	0,74	0,7	0,7	0,64	0,65	0,65	0,73
P1	-	0,77	0,73	0,76	0,77	0,74	0,72	0,71	0,72	0,77
P2		-	0,79	0,79	0,81	0,78	0,79	0,77	0,85	0,82
P3			-	0,78	0,77	0,78	0,73	0,76	0,76	0,82
P4				-	0,81	0,8	0,78	0,79	0,8	0,81
P5					-	0,77	0,73	0,77	0,78	0,83
P6						-	0,76	0,75	0,77	0,79
P7							-	0,71	0,75	0,79
P8								-	0,71	0,76
P9									-	0,76
P10										-

Il ressort de ce tableau que l'utilisation d'une phrase seule ne permet pas de bonne corrélation avec l'extrait de texte. Cependant, des valeurs de corrélation égales ou supérieures à 0,8 apparaissent, lorsque certaines phrases sont appariées. La corrélation la plus forte (0,85) est obtenue avec la combinaison des phrases P2 et P9 (15 mots). Lorsque la phrase P4 est appariée, elle obtient quatre bonnes corrélations, de même pour la phrase P10. Ces deux phrases combinées ensemble, soit 18 mots, donnent une corrélation de 0,81. Nous notons également que la combinaison des phrases P7 et P8 avec d'autres phrases a systématiquement une valeur de corrélation inférieure à 0,8. Ces deux phrases étant celles contenant le moins de mots (6 mots chacune).

Ces premiers résultats sont encourageants sachant que le paragraphe du texte de « La chèvre de Monsieur Seguin » lu par les locuteurs comprend entre 47 et 53 mots et entre 64 et 71 syllabes (selon la version demandée au patient lors de l'enregistrement). Mais il ressort clairement que l'utilisation d'une phrase seule paraît illusoire dans l'estimation et le suivi du niveau de sévérité d'un patient. Ces premiers résultats indiquent aussi qu'une combinatoire de phrases pourrait conduire à des résultats satisfaisants et ce tout en allégeant la tâche des locuteurs.

Le tableau 4 ci-dessous montre les valeurs des facteurs de corrélation obtenues pour des combinaisons de 3 phrases. Chacune des phrases (de P1 à P10) a été combinée avec les couples de phrases mentionnées ci-dessus. P4P10 et P2P9 car ces combinaisons ont obtenu de bons scores de corrélation. Nous avons également voulu combiner le couple P7P8, ayant obtenu les

scores de corrélation les plus bas, pour voir l'incidence d'une combinaison avec une troisième phrase.

Tableau 4 : Corrélation entre le texte de « La chèvre de Monsieur Seguin » et l'association de 3 phrases de Combescure :

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P4P10	0,81	0,84	0,84	-	0,85	0,83	0,83	0,82	0,82	-
P2P9	0,79	-	0,82	0,83	0,84	0,82	0,82	0,79	-	0,83
P7P8	0,74	0,79	0,77	0,79	0,76	0,77	-	-	0,76	0,79

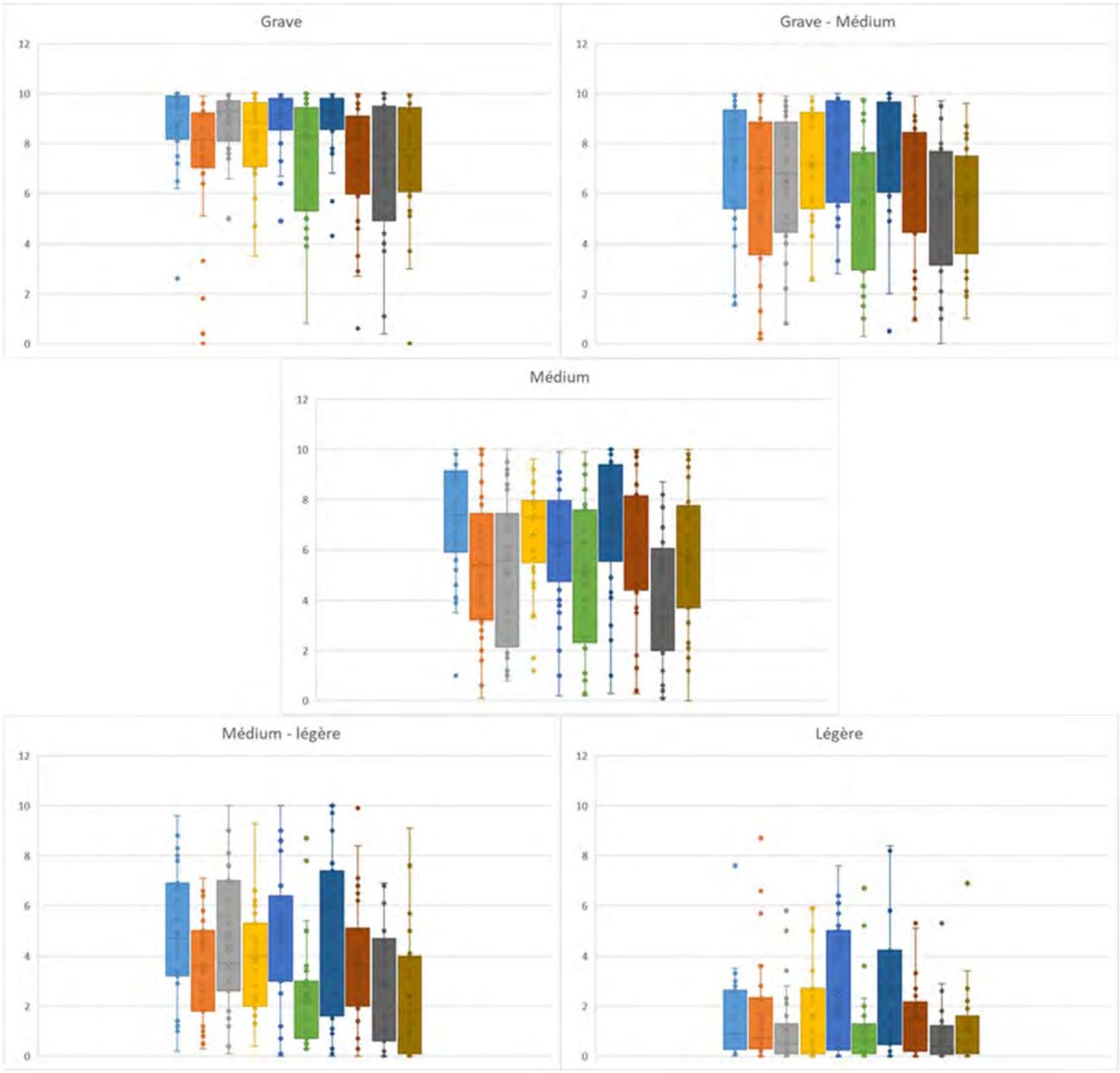
Nous observons que de bons facteurs de corrélation sont obtenus en combinant n'importe quelles phrases de P1 à P10 avec les couples de phrases P4P10 et P2P9, soit respectivement une fourchette de mots compris entre 24 mots et 26 mots, et 21 mots et 23 mots, et une fourchette de syllabes comprises entre 35 et 41, et entre 28 et 36 syllabes.

En ce qui concerne les combinaisons des différentes phrases avec le couple P7P8 (nombre de mots compris entre 18 mots et 21 mots, pour 27 à 32 syllabes), elles obtiennent des scores insuffisants même si une légère augmentation des facteurs de corrélation est observée.

Il ressort de ces résultats que la tâche des patients pourrait être allégée dans la mesure où un nombre réduit de mots (23-24 mots) et/ou de syllabes (35-41 syllabes) serait utilisé.

Les phrases de Combescure n'étant pas équivalentes en termes de mots/syllabes, nous nous sommes interrogés sur une possible incidence quant à leur notation respective par les auditeurs. La figure suivante indique par niveau de gravité, la représentation en box plots de la distribution des scores par phrases, tout auditeur confondu.

Figure 6 : Distribution des scores par type de phrases, de P1 à P10, et niveau de sévérité, tout auditeur confondu



Il paraît difficile de tirer des tendances générales de ces résultats, même s'il apparaît clairement que ces phrases sont perçues différemment par les auditeurs et impacte leur notation. Par exemple, si nous considérons les phrases P9 et P10 dans les cas de sévérité « grave », nous notons une tendance de sous-évaluation de certains auditeurs (médiane inférieure à 8). De même, nous pouvons noter que, d'une façon générale, la phrase P7 conduit à un niveau de sévérité supérieur aux autres phrases. Nous constatons que la phrase P7 est une phrase

interrogative qui suggère une certaine prosodie. Cependant, il est difficile de s'avancer sur l'influence de celle-ci dans la notation, dans le cadre de ce mémoire. Aucune consigne n'étant donnée au patient lors de la tâche, celui-ci a pu appliquer la prosodie qu'il souhaitait sur toutes les phrases, pouvant peut-être influencer l'auditeur.

3.3. Traitement automatique

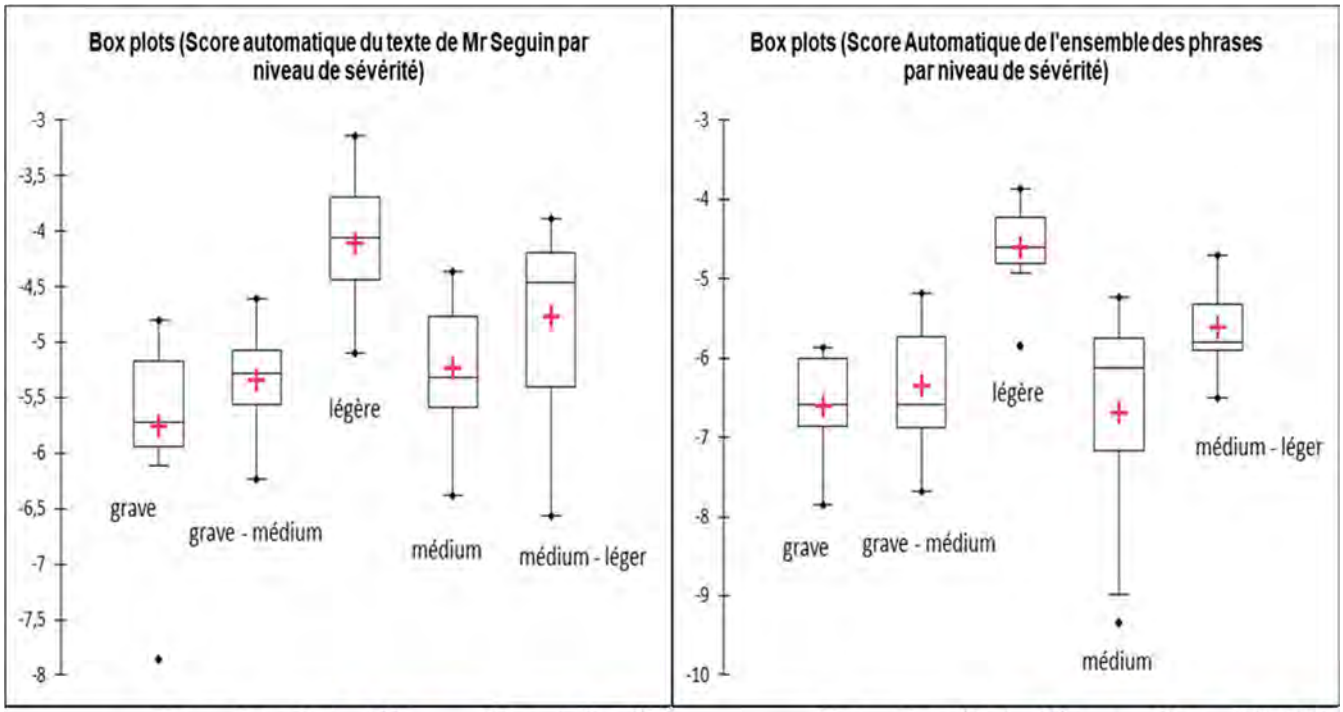
Une analyse automatique des enregistrements a été réalisée. Les scores obtenus par cette analyse correspondent à la moyenne des scores de vraisemblance des phonèmes du texte et pour chacune des phrases. La valeur du score est toujours inférieure à 0, cette valeur indiquant une prononciation parfaite au regard du modèle générique. Plus la prononciation est décalée par rapport au modèle de référence, plus la valeur est négative. Les scores obtenus par les analyses automatique et perceptive variant dans des sens opposés, les valeurs des facteurs de corrélation sont négatives.

Par ailleurs, la valeur du score automatique n'étant pas bornée, c'est-à-dire qu'elle ne tend pas vers une valeur minimale seuil, cela implique qu'il faut toujours raisonner en termes de corrélation avec le score perceptif. De plus, les scores perceptifs sont bornés ce qui peut induire, pour les cas extrêmes de sévérité, une détérioration des facteurs de corrélation. Par exemple, si tous les locuteurs de niveau de sévérité « grave » avaient la note de 10, les scores automatiques se positionneraient sur une ligne horizontale, la corrélation étant alors nulle.

3.3.1. Analyse des scores automatiques en fonction des niveaux de sévérité

La figure 7 ci-dessous présente la distribution des scores automatiques en fonction des niveaux de sévérité sous forme de box plots. Les tableaux situés en-dessous indiquent si les différences sont significatives, ou non (test de Mann-Whitney).

Figure 7 : Scores automatiques en fonction des niveaux de sévérité



	grave	grave - médium	légère	médium	médium - léger
grave	Non	Non	**Oui**	Non	**Oui**
grave - médium	Non	Non	**Oui**	Non	Non
légère	**Oui**	**Oui**	Non	**Oui**	Non
médium	Non	Non	**Oui**	Non	Non
médium - léger	**Oui**	Non	Non	Non	Non

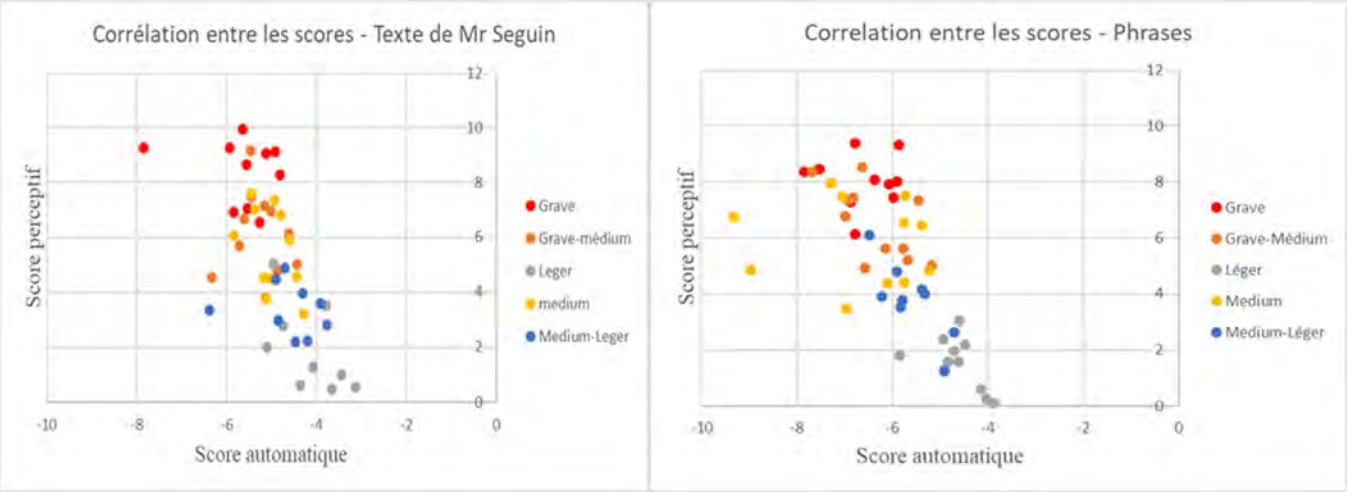
	grave	grave - médium	légère	médium	médium - léger
grave	Non	Non	**Oui**	Non	**Oui**
grave - médium	Non	Non	**Oui**	Non	Non
légère	**Oui**	**Oui**	Non	**Oui**	**Oui**
médium	Non	Non	**Oui**	Non	Non
médium - léger	**Oui**	Non	**Oui**	Non	Non

Nous observons que les scores automatiques permettent d'individualiser le niveau de sévérité « légère » des autres niveaux. Ce constat se vérifie indifféremment pour l'extrait de texte de « La chèvre de Monsieur Seguin » et l'ensemble des phrases de Combescure. Nous notons également une différence significative entre les niveaux de sévérité « médium-légère » et « grave ».

3.3.2. Corrélation entre les scores automatiques et les scores perceptifs

La confrontation des moyennes des scores automatiques et perceptifs en fonction des niveaux de sévérité est présentée dans la figure 8. L'objectif est de visualiser, ou pas, si une cohérence se dessine entre les scores automatiques (abscisses) et les scores perceptifs (ordonnées) obtenus d'une part pour le texte et d'autre part pour les 10 phrases de Combescure.

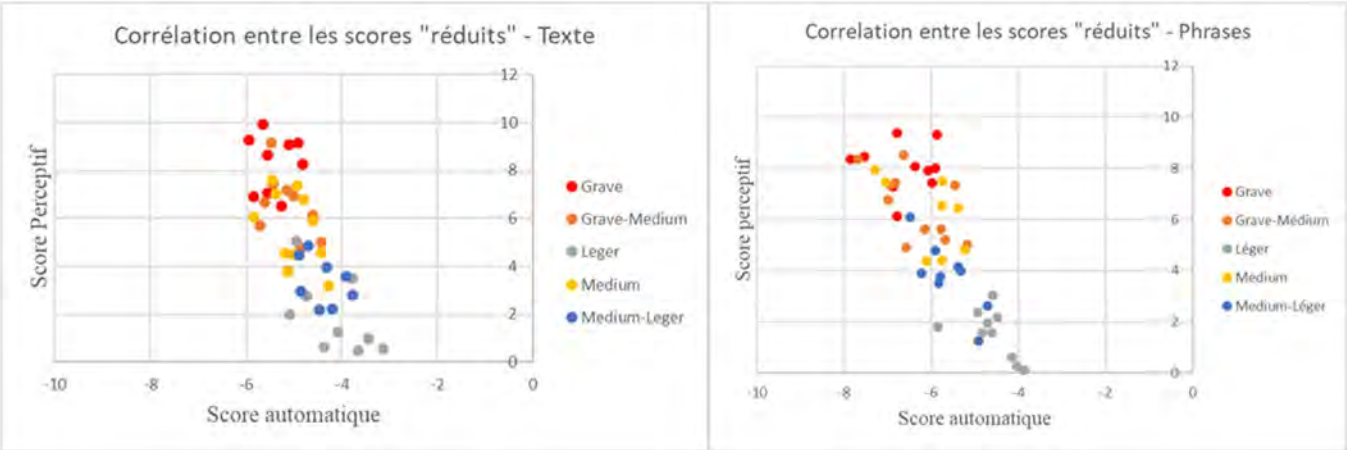
Figure 8 : Corrélation entre les scores automatique et perceptif par niveau de sévérité



Nous pouvons constater que les niveaux de sévérité sont regroupés en nuages de points présentant des aires de chevauchement entre niveaux voisins. Comme attendu, le nuage de points correspondant au niveau de sévérité « Grave » est regroupé en haut à gauche et le nuage de points pour le niveau de sévérité « Légère » en bas à droite.

Nous constatons par ailleurs que dans les deux cas, trois points s'écartent de la tendance générale. Si nous éliminons ces points, la figure suivante est obtenue (Figure 9).

Figure 9 : Corrélation entre les scores automatique et perceptif « réduits » par niveau de sévérité



Les valeurs des facteurs de corrélation calculées pour le texte de Monsieur Seguin, sont de -0,64 ou de -0,75 selon que nous prenions en compte respectivement, l'ensemble des scores ou en éliminant les trois valeurs qui s'écartent du nuage de points.

En ce qui concerne les phrases de Combescure, ces scores passent respectivement de -0,67 à -0,81.

Les trois points, dégradant la corrélation entre les scores de l'analyse automatique et l'analyse perceptive de l'extrait de texte, correspondent à trois locuteurs de niveau de sévérité différent : un locuteur est du niveau « grave » (score automatique/perceptif de -7,8/9,2), un locuteur du niveau « grave-médium » (score automatique/perceptif de -6,3/4,5) et un locuteur du niveau « médium-légère » (score automatique/perceptif de -6,6/4,5). La moyenne des scores automatiques est de -5,65 pour le niveau de sévérité « grave » et les écarts à cette moyenne sont tous compris entre 0,8 et -0,2, à l'exception du locuteur concerné dont l'écart est de -2,2. Pour la moyenne des scores automatiques « grave-médium », la moyenne est de -5,25, les écarts étant tous compris entre 0,8 et -0,5, alors que l'écart du locuteur impliqué est de -1,1. Enfin, la moyenne des scores automatiques « médium-légère » est de -4,6, les écarts se situant entre -0,28 et 0,8 alors que le locuteur concerné présente un écart de -1,8.

Concernant les 3 points dégradant la corrélation entre les scores de l'analyse automatique et de l'analyse perceptive des phrases de Combescure, ils correspondent à trois locuteurs de niveau de sévérité « médium ». Pour les deux locuteurs les plus décalés sur la partie gauche du nuage de points, nous constatons que les trois mêmes phrases (à savoir P2-P5-P9) ont une moyenne des scores de vraisemblance (au regard du modèle générique) inférieure à -10. Pour le dernier locuteur concerné, seule la phrase P2 a obtenu un score inférieur à -10.

Nous pensons que ces observations illustrent assez bien les différences d'évaluation entre une méthode d'analyse intégrative (l'analyse perceptive) et ce d'autant que les auditeurs « naïfs » devaient estimer un niveau global de sévérité du signal, et une analyse automatique qui porte sur l'identification de phonèmes par rapport à un modèle.

Cependant, les scores obtenus au travers de l'analyse automatique suggèrent que les niveaux de sévérité peuvent être évalués indifféremment par l'utilisation du texte de « La chèvre de Monsieur Seguin » ou par les 10 phrases de Combescure. Dans la perspective d'une réduction de la tâche du patient, il convient maintenant d'estimer dans quelle mesure nous pouvons réduire le nombre de phrases, sachant que le texte comprend 47 à 53 mots (64 à 71 syllabes) et que l'ensemble des phrases en compte 76 mots (110 syllabes).

3.3.3. Estimation des facteurs de corrélation selon les combinaisons de phrases

Si nous admettons que les scores obtenus par analyse automatique sont globalement bien corrélés avec ceux de l'analyse perceptive, il est légitime de confronter les scores automatiques obtenus pour l'ensemble des phrases ou de combinaisons de phrases avec ceux obtenus pour le texte, et ce, dans une perspective d'estimer s'il est possible, ou pas, de réduire le nombre de phrases.

Le facteur de corrélation entre les scores automatiques de l'ensemble des phrases et celui du texte est de 0,64. Le tableau suivant indique les facteurs de corrélation entre score automatique de chaque phrase, de P1 à P10 (première ligne du tableau), puis les combinaisons de phrases (2 à 2), avec le score automatique du texte.

Tableau 5 : Corrélation entre le texte de « La chèvre de Monsieur Seguin » et les phrases de Combescure en score automatique :

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Phrase seule	0,47	0,35	0,65	0,65	0,58	0,47	0,61	0,48	0,37	0,48
P1	-	0,46	0,61	0,63	0,60	0,55	0,62	0,55	0,49	0,53
P2		-	0,53	0,50	0,49	0,45	0,50	0,43	0,39	0,45
P3			-	0,72	0,69	0,64	0,70	0,66	0,57	0,64
P4				-	0,69	0,61	0,69	0,64	0,54	0,64
P5					-	0,60	0,66	0,61	0,53	0,60
P6						-	0,59	0,54	0,45	0,54
P7							-	0,65	0,54	0,64
P8								-	0,46	0,56
P9									-	0,48
P10										-

Les facteurs de corrélation obtenus par phrase sont compris entre 0,35 et 0,65 (seulement 30% des valeurs sont ≥ à 0,6). La valeur de ce facteur de corrélation calculée par combinaisons de phrases est comprise entre 0,39 et 0,72 avec 44% des valeurs ≥ à 0,6. Ce premier résultat, indique comme précédemment, que la combinaison de phrases améliore la corrélation mais reste insuffisante.

Au regard de ces résultats, d'autres combinaisons comprenant un nombre plus important de phrases ont été réalisées, et les facteurs de corrélation calculés.

Tableau 6 : Corrélation entre le texte de « La chèvre de Monsieur Seguin » et l'association de 3 phrases et plus, en score automatique :

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P4P10	0,63	0,54	0,70	-	0,67	0,63	0,70	0,65	0,57	-
P3P4	0,68	0,60	-	-	0,73	0,69	0,73	0,70	0,64	0,70
P3P7	0,67	0,60	-	0,73	0,72	0,68	-	0,70	0,64	0,70
P3P4P5	0,70	0,63	-	-	-	0,71	0,74	0,72	0,67	0,71
P3P4P7	0,71	0,64	-	-	0,74	0,71	-	0,72	0,67	0,73
P3P4P5P7	0,72	0,66	-	-	-	0,72	-	0,73	0,69	0,73

Le choix des combinaisons à effectuer au regard des bons scores obtenus par certaines combinaisons de deux phrases (cf. Tableau 5). Des combinaisons de plus de 3 phrases ont été réalisées pour estimer si les corrélations étaient meilleures. D'un part, nous notons seulement deux valeurs < à 0,60 avec la combinaison de phrases P4P10. D'autre part, nous observons un plafonnement des valeurs des facteurs de corrélation avec des combinaisons de 4 et 5 phrases.

Le tableau ci-dessous récapitule le nombre de mots (et le nombre de syllabes) pour chaque combinaison réalisée.

Tableau 7 : Nombre de mots et nombre de syllabes (indiqué entre parenthèses) pour les combinaisons de phrases :

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P4P10	26 (41)	25 (36)	26 (37)	-	26 (37)	25 (38)	24 (35)	24 (37)	26 (38)	-
P3P4	26 (40)	25 (35)	-	-	26 (36)	25 (37)	24 (34)	24 (36)	26 (37)	26 (37)
P3P7	22 (32)	21 (27)	-	24 (34)	22 (28)	21 (29)	-	20 (28)	22 (29)	22 (29)
P3P4P5	34 (50)	33 (45)	-	-	-	33 (47)	32 (44)	32 (46)	34 (47)	34 (47)
P3P4P7	32 (48)	31 (43)	-	-	32 (44)	31 (45)	-	30 (44)	32 (45)	32 (45)
P3P4P5P7	40 (58)	39 (53)	-	-	-	39 (55)	-	38 (54)	40 (55)	40 (55)

Comme indiqué précédemment, 6 locuteurs obtiennent des scores, en analyse automatique, différents de ceux obtenus au sein du niveau de sévérité auquel ils appartiennent. Nous nous sommes posé la question de savoir ce qui les différencie en particulier des autres locuteurs. Nous avons constaté que le seul critère qui les différencie concerne le lieu d'enregistrement. Les enregistrements de ces 6 locuteurs ont tous été réalisés au sein de l'Hôpital Larrey, pour rappel dans une pièce non insonorisée avec un casque-micro. Il reste à déterminer les causes précises ayant pu générer de tels écarts. En effet, la qualité de

l'enregistrement, le bruit, l'écho et la réverbération ainsi que les pauses dans la parole ont pu engendrer des erreurs dans le système d'alignement automatique, lors de la détection des frontières des différents phonèmes. Ces erreurs peuvent alors affecter les scores de vraisemblances qui seront très bas. Nous notons que les 6 locuteurs présentent les scores automatiques les plus sévères.

Ainsi, cet élément factuel nous a conduits à calculer à nouveau les facteurs de corrélation (pour chacune des phrases et combinaisons de phrases réalisées auparavant) selon que l'on écarte les 6 locuteurs, les 3 locuteurs identifiés via le texte ou les 3 locuteurs identifiés via les phrases. Le tableau ci-dessous présente les valeurs des facteurs de corrélation obtenus en écartant les 3 « locuteurs phrases ». Afin de faciliter les comparaisons, les précédents tableaux sont rappelés. Ces résultats montrent, dans tous les cas, une amélioration des facteurs de corrélation. Ils mettent en relief l'importance probable et la standardisation des conditions d'enregistrement sur les résultats de l'analyse automatique et soulignent les soins à y apporter.

Tableau 8 : Confrontation des tableaux des facteurs de corrélation des combinaisons de phrases lors de la suppression des « locuteurs phrases » :

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Phrase seule	0,46	0,39	0,66	0,65	0,68	0,46	0,63	0,50	0,37	0,60
P1	-	0,50	0,61	0,63	0,66	0,55	0,62	0,56	0,50	0,58
P2		-	0,60	0,54	0,61	0,49	0,57	0,49	0,43	0,57
P3			-	0,73	0,75	0,66	0,72	0,67	0,60	0,73
P4				-	0,75	0,62	0,70	0,65	0,56	0,71
P5					-	0,65	0,73	0,69	0,60	0,75
P6						-	0,61	0,54	0,46	0,60
P7							-	0,65	0,59	0,74
P8								-	0,46	0,67
P9									-	0,54
P10										-

Moins les 3 « locuteurs phrases »

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
Phrase seule	0,47	0,35	0,65	0,65	0,58	0,47	0,61	0,48	0,37	0,48
P1	-	0,46	0,61	0,63	0,60	0,55	0,62	0,55	0,49	0,53
P2		-	0,53	0,50	0,49	0,45	0,50	0,43	0,39	0,45
P3			-	0,72	0,69	0,64	0,70	0,66	0,57	0,64
P4				-	0,69	0,61	0,69	0,64	0,54	0,64
P5					-	0,60	0,66	0,61	0,53	0,60
P6						-	0,59	0,54	0,45	0,54
P7							-	0,65	0,54	0,64
P8								-	0,46	0,56
P9									-	0,48
P10										-

Tous les locuteurs compris

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P4P10	0,67	0,63	0,76	-	0,78	0,67	0,76	0,72	0,63	-
P3P4	0,69	0,65	-	-	0,77	0,70	0,75	0,71	0,66	0,76
P3P7	0,68	0,67	-	0,75	0,77	0,70	-	0,72	0,68	0,78
P3P4P5	0,74	0,71	-	-	-	0,75	0,78	0,76	0,72	0,80
P3P4P7	0,72	0,69	-	-	0,78	0,72	-	0,74	0,71	0,78
P3P4P5P7	0,76	0,73	-	-	-	0,76	-	0,77	0,74	0,80

Moins les 3 « locuteurs phrases »

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
P4P10	0,63	0,54	0,70	-	0,67	0,63	0,70	0,65	0,57	-
P3P4	0,68	0,60	-	-	0,73	0,69	0,73	0,70	0,64	0,70
P3P7	0,67	0,60	-	0,73	0,72	0,68	-	0,70	0,64	0,70
P3P4P5	0,70	0,63	-	-	-	0,71	0,74	0,72	0,67	0,71
P3P4P7	0,71	0,64	-	-	0,74	0,71	-	0,72	0,67	0,73
P3P4P5P7	0,72	0,66	-	-	-	0,72	-	0,73	0,69	0,73

Tous les locuteurs compris

0,60≥Valeur>0,69 0,70≥Valeur>0,79 Valeur≥0,80 Valeur≤ à la valeur avec tous les locuteurs compris

4. DISCUSSION

Les résultats obtenus suggèrent que, globalement, analyse perceptive et analyse automatique convergent quant à l'évaluation du niveau de sévérité. L'extrait de texte de « La chèvre de Monsieur Seguin » pourrait être substitué par les phrases de Combescure. Par ailleurs, il serait fortement envisageable de réduire la tâche du patient afin de mener à bien une évaluation du niveau de sévérité d'altération de la parole. Dans ce mémoire, les tâches qui ont été demandées au patient lors de leur suivi comprennent le premier paragraphe du texte de « La chèvre de Monsieur Seguin » d'Alphonse Daudet (comprenant entre 47 et 53 mots, selon l'extrait soumis au patient) et les 10 phrases de la première liste de Combescure (comprenant 76 mots), pour un total de plus de 130 mots. Une évaluation pourrait être menée avec 25-30 mots.

Ces résultats, aussi encourageants soient-ils, doivent être pondérés. En effet, notre étude n'a porté que sur 51 patients atteints d'un cancer des VADS. Par ailleurs, chaque locuteur n'a été évalué qu'au travers de 33 auditions (30 auditions de phrases et 3 auditions de texte). L'augmentation des effectifs et du volume de données permettrait de confirmer ou d'infirmer statistiquement ces résultats. Il serait intéressant de déterminer l'importance relative du nombre de mots versus le nombre de syllabes sur les notations. Enfin, des pistes d'amélioration peuvent être proposées au regard des difficultés et questionnements rencontrés au cours de l'analyse des données.

Le problème des niveaux de sévérité

L'estimation des niveaux de sévérité au travers d'une analyse montre certaines difficultés et limites. Une des difficultés porte sur la définition même des groupes de sévérité. Nous pouvons nous interroger sur le bien-fondé de distinguer 5 groupes de sévérité, groupes censés être des espaces bornés mais dont les bornes ne sont pas clairement définies. Cela se traduit donc par une dispersion moindre des scores obtenus au sein des groupes extrêmes « grave » et « légère » comparativement aux groupes intermédiaires. Autrement dit, les cas tranchés constituent des évidences et sont faciles à classer dans une catégorie. Les cas intermédiaires sont plus difficiles à positionner dans une catégorie donnée, d'autant plus lorsque le nombre de catégories est grand et qu'elles ne sont pas définies de façon explicite. Cette difficulté n'est pas propre à cette étude et a été soulignée par ailleurs dans le cadre d'analyse perceptive catégorielle (Ghio et al., 2014). La problématique des groupes de sévérité n'est pas

forcément dans leur nombre mais plutôt dans leur définition (leur contenu et leurs bornes), en étroite relation avec la question (et les consignes associées) posée aux auditeurs ainsi que le statut de ces mêmes auditeurs.

Jury expert ou jury naïf ?

Dans la littérature, différentes compositions de jury ont été utilisées ou préconisées. Cette composition porte sur, d'une part, la qualité des auditeurs (jurys composés de membres experts, de membres naïfs ou de membres naïfs ayant suivi un apprentissage intensif) ; d'autre part, leur composition en nombre (Ghio et al., 2014).

Dans un contexte clinique, l'estimation peut être réalisée par une équipe de thérapeutes, personnels fortement impliqués dans le processus d'amélioration. Concernant ces professionnels, il a été relevé un effet d'apprentissage conduisant à un risque de surévaluation du niveau d'intelligibilité des locuteurs. Les jurys composés de membres naïfs conduisent à une plus grande variabilité des scores. Cette variabilité des scores était d'autant plus prononcée dans les catégories médianes de sévérité. Afin de réduire cette variabilité des résultats obtenus par le biais d'auditeurs naïfs, (Ghio et al., 2014) ont mené une expérimentation préconisant d'entraîner les auditeurs. Leurs résultats ont montré une amélioration de la notation en termes de convergence et, principalement dans les catégories intermédiaires. Ce même auteur considère que l'analyse perceptive reste la référence de l'évaluation de sévérité, dans la mesure où elle est conduite par des membres experts, pour atteindre ainsi une fiabilité raisonnable de la perception. L'appel à des experts réduit le poids de la subjectivité génératrice de variabilité. L'inconvénient majeur de ce type de procédure consiste dans leur temps (composition du jury, temps de mobilisation des experts, temps d'analyse, …) et les coûts globaux de mise en œuvre.

Dans le cadre de notre étude, le jury était composé de 30 membres auditeurs naïfs. Les scores obtenus, sont très variables d'un auditeur à l'autre. Cette observation est en accord avec de précédentes études. En effet, les auditeurs naïfs n'ont pas de norme spécifique pour juger de la qualité de voix pathologiques. Leur jugement repose donc sur leur propre norme de ce qu'il considère comme une voix normale. Chaque auditeur ayant son propre référentiel, il en résulte une grande variabilité des résultats (Kreiman et al., 1993). Par ailleurs, les notations attribuées à une même phrase par un même auditeur pourraient être influencées selon que son écoute est précédée d'un locuteur de faible intelligibilité ou au contraire sans défaut majeur. Quoi qu'il en soit, l'ensemble de ces observations met en relief la complexité de tests faisant intervenir un jury d'écoute, de façon plus prononcée dans le cas d'auditeurs naïfs, du fait de la grande

variabilité individuelle des processus de perception. Cependant, une convergence des résultats a été reportée dans le cadre d'une étude comparant ceux d'auditeurs experts à ceux d'auditeurs naïfs (Roux et al., 2016). Cette convergence observée est très bonne dans le cas de questions simples débouchant sur une réponse de type binaire (présence/absence). Les scores divergent d'autant en fonction de la « professionnalisation » de la question posée aux auditeurs.

Compte tenu des éléments indiqués ci-dessus, un jury expert aurait-il conduit à des résultats profondément différents ? En termes de variabilité, nous pouvons émettre l'hypothèse que celle-ci serait meilleure. Il serait intéressant de soumettre à un jury d'expert le même panel de locuteurs afin de confirmer ou d'infirmer cette hypothèse.

Par ailleurs, il est légitime de se poser la question suivante : les auditeurs évaluent-ils la même chose ? Tous les auteurs s'accordent sur la nécessité de faire appel à des jurys homogènes dans leur constitution (Crevier-Buchman, 2012 ; Tribout, 2013). Le choix de faire appel à des auditeurs naïfs vise à confronter la parole du patient à des auditeurs qu'il est susceptible de rencontrer dans sa vie quotidienne. La réalisation d'un tel test a une visée écologique débouchant sur une estimation de l'impact social de l'altération et de son suivi dans le temps. Pour des cliniciens, leur niveau d'expertise leur permettra une évaluation plus pointue dans l'identification des altérations. Leur évaluation est la plus à même de déboucher sur la mise en place d'un protocole correctif. Les évaluations réalisées par les deux types de jurys sont complémentaires.

Peut-on améliorer la convergence des scores d'un jury naïf ?

Il a été reporté par (Ghio et al., 2014) que l'apprentissage de membres d'un jury naïf améliore la convergence des résultats des catégories de sévérité médiane. Dans notre cas, il aurait été intéressant de définir un protocole plus précis concernant le déroulement des auditions. Nous pourrions tester deux pistes d'amélioration : i) effectuer les auditions dans un environnement contrôlé et, ii) donner des consignes améliorées.

i) Nous n'avons pas d'information concernant le contexte dans lequel les auditions se sont déroulées (matériel d'écoute, environnement sonore, …) mais nous pouvons penser qu'il était très variable. Faire passer les auditions dans des conditions équivalentes pour tous les auditeurs devrait réduire la variabilité des résultats. D'autre part, nous pouvons supposer une plus grande concentration des auditeurs, leur attention ne pouvant pas être mobilisée ailleurs.

ii) La consigne fournie avec les enregistrements était : « vous allez écouter des enregistrements de voix, il vous est demandé après chaque écoute de placer sur une droite un trait marquant le degré d'altération globale du signal sonore […] ». Cette consigne laisse trop de degré de liberté à l'auditeur concernant sa méthode de travail. Nous n'avons pas d'information sur le nombre d'écoutes réalisées pour une même phrase, si une première écoute globale a été réalisée avant de procéder à la notation, … Nous pouvons raisonnablement penser que différentes méthodes de travail ont été utilisées et ont contribué à la variabilité des scores inter auditeurs. La plus faible convergence des scores concernant les groupes intermédiaires, il serait intéressant de tester dans un prochain travail, un protocole en deux temps. Lors d'une première écoute, l'auditeur positionnerait les enregistrements dans deux catégories de part et d'autre de la valeur médiane de l'échelle. La seconde écoute permettrait d'affiner sa notation. L'avantage serait de disposer de scores réalisés dans un contexte et un protocole homogènes.

Peut-on alléger le nombre de tâches des locuteurs ?

Les avantages seraient multiples. Pour le patient, cela limiterait, d'une part la fatigue observée lors de tests de longue durée et, d'autre part, leur temps de mobilisation à réaliser cette tâche. Pour les structures en charge de la réalisation, les gains de temps de récolte des enregistrements ainsi que d'analyse, seraient appréciables.

Les résultats obtenus dans le cadre de ce travail sont encourageants. D'une part, des résultats similaires ont été obtenus, que nous utilisions un extrait de texte ou les 10 phrases de Combescure. D'autre part, une réduction de moitié (en termes de mots) serait envisageable. De façon intéressante, il a déjà été reporté par (Fontan, 2012) que :

- i) le score d'intelligibilité est croissant selon que le locuteur prononce des mots, des phrases ou un discours.

- ii) de manière plus précise, les scores d'intelligibilité sont influencés par le nombre de mots contenus dans une phrase. Plus la longueur de la phrase était importante, plus le score d'intelligibilité augmentait.

- iii) enfin, une variation significative de 5% des scores est observée entre des phrases courtes (5-9 mots) et des phrases longues (10-15 mots).

Dans le cadre de ce mémoire, l'influence de la prosodie sur les scores d'intelligibilité n'a pas pu être prise en compte. (Nocaudie et al., 2018) ont montré que les résultats entre groupe témoins et groupe patients sont peu tranchés pour certaines tâches (tâche de syntaxe et de

modalité). Seule la tâche de focus donne des résultats significativement différents entre les deux groupes.

Des travaux ultérieurs sont donc à mettre en place en se focalisant sur le nombre de mots ou de syllabes et de valider ou d'infirmer nos résultats. De plus, il conviendrait de s'assurer de l'équilibre phonétique des quelques phrases retenues.

L'analyse automatique : une aide à l'analyse des données

Une précédente corrélation par analyse automatique avait été obtenue, pour le texte de Monsieur Seguin, entre les scores automatiques et perceptifs, lors d'une précédente étude dans le cadre du projet C2SI (Fredouille, 2019). Ce facteur de corrélation (0,89) est supérieur à celui (0,64) que nous avons obtenu dans le cadre de ce travail. Plusieurs raisons pourraient concourir à cette différence :

i) Les conditions d'enregistrements n'ont pas été réalisées dans des conditions équivalentes selon les différents établissements de passation. Les personnes ayant une expertise dans l'analyse automatique considère que la qualité des enregistrements semble de moindre qualité comparativement au corpus du C2SI.

ii) Les effectifs de locuteurs sont bien inférieurs dans notre étude comparativement au projet C2SI. Respectivement, 51 locuteurs pour 87 patients confrontés à 42 sujets témoins.

iii) Enfin, il faut garder en mémoire, que dans notre étude, l'extrait de texte était plus ou moins long selon les locuteurs, ce qui a probablement impacté les résultats.

Il n'en reste pas moins que les résultats de l'analyse par traitement automatique du texte et des phrases sont convergents, autrement dit, les phrases de Combescure et le texte de Monsieur Seguin pourraient être utilisés indifféremment pour évaluer le niveau de sévérité de l'intelligibilité d'un patient.

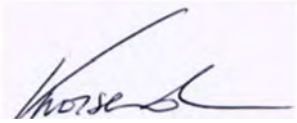
Peut-on substituer l'analyse perceptive par l'analyse automatique ? Les auditeurs des jurys, indépendamment de leur niveau d'expertise, analysent la voix d'un locuteur en s'appuyant sur la puissance informative de son écoute analytique, au travers de l'oreille humaine en tant que récepteur. Cette dernière fonctionne de manière intégrative et ne permet pas toujours de distinguer et encore moins de quantifier des paramètres fréquentiels et énergétiques. L'analyse de certains de ces paramètres a montré une bonne corrélation avec les niveaux d'intelligibilité de patients atteints de cancer des VADS (Sicard et al., 2017). L'un des avantages de cette analyse est qu'elle met en correspondance des données acoustiques,

articulatoires, comportementales, praxiques et de qualité de vie. Il ressort de ces considérations qu'on ne mesure pas les mêmes éléments. En effet, nous constatons que selon le type d'analyse effectuée, les facteurs de corrélation obtenus sur les phrases et les combinaisons ne sont pas équivalents. Par exemple, la combinaison de phrases P2P9 qui obtient la meilleure corrélation en analyse perceptive, obtient au contraire un score moins favorable par analyse automatique. De même, la phrase P7, dont la prosodie a probablement influencé défavorablement son score en analyse perceptive, obtient globalement un bon score en analyse automatique.

5. CONCLUSION

Pour conclure, nous pensons que si l'analyse perceptive est une analyse sujette à la subjectivité liée à l'auditeur, elle a l'avantage d'être une analyse intégrative et écologique. Parallèlement, l'analyse automatique, analyse de type mécanistique, fournit des résultats objectifs par comparaison entre la voix d'un patient et un modèle pré-établi. Les résultats obtenus par les deux types d'analyse sont complémentaires, une analyse ne remplaçant pas l'autre. Dans les deux cas, les résultats obtenus sont encourageants dans une perspective de réduction de la tâche de lecture des locuteurs. Les deux méthodes suggèrent que l'évaluation pourrait être menée sur une/des phrases de 25-30 mots. Ces résultats se doivent d'être confirmés et une réflexion conduite afin de bien définir le contenu des phrases.

Virginie Woisard, le 23 août 2019



BIBLIOGRAPHIE

Astésano, C., Balaguer, M., Farinas, J., Fredouille, C., Gaillard, P., Ghio, A., … Woisard, V. (2018). Carcinologic Speech Severity Index Project: A Database of Speech Disorder Productions to Assess Quality of Life Related to Speech After Cancer. *Language Resources and Evaluation Conference (LREC)*. Présenté à Miyazaki, Japan. https://hal.archives-ouvertes.fr/hal-01770168

Balaguer-Navarro, M. (2018). *Construction d'un score Carcinologic Speech Severity Index (C2SI) automatique* [Mémoire Master 2 - Recherche Epidémiologie clinique]. Toulouse: IRIT (Institute de Recherche en Informatique de Toulouse) - Université Paul Sabatier Toulouse III - Faculté de Médecine Purpan

Brin, F., Courrier, C., Lederlé, E. (2018). *Dictionnaire d'orthophonie* (4ème éd.). Isbergues, France: Ortho Edition

Crevier-Buchman, L. (2012). *Contribution à la compréhension de la voix et de la parole normale et pathologique* [Mémoire d'habilitation à diriger des recherches – Phonétique clinique]. Aix-Marseille Université.

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

Fontan, L. (2012). *De la mesure de l'intelligibilité à l'évaluation de la compréhension de la parole pathologique en situation de communication* [Thèse – Linguistique]. Toulouse - Université Toulouse II Le Mirail. https://tel.archives-ouvertes.fr/tel-00797883

Fredouille, C. (2019). *Traitement automatique appliqué aux troubles de la parole : approches, résultats et limites - Atelier « Ressources et outils de traitement automatique pour la pratique clinique ainsi que la recherche en parole atypique et pathologique »*. Présentée aux Journées de Phonétique Clinique, Mons, Belgique.

Gaillard, P., Billieres, M., Magnen, C. (2007). La surdité phonologique illustrée par une étude de catégorisation des voyelles françaises perçues par les hispanophones. In: Proc. Percepción y Realidad., Valladolid, Espagne ; pp. 187-196

Ghio, A., Dufour, S., Pouchoulin, G., Revis, J., Robert, D., et al. (2014). Contributions expérimentales à l'élaboration d'un protocole robuste d'évaluation perceptive des troubles de la voix et de la parole. P*arole*, pp.85-101. https://hal.archives-ouvertes.fr/hal-01294774

Ghio, A., Giusti, L., Blanc, E., Pinto, S., Lalain, M., Robert, D., … Woisard, V. (2016). Quels tests d'intelligibilité pour évaluer les troubles de production de la parole ? *Journées d'Etude sur la Parole*, 589-596. https://hal.archives-ouvertes.fr/hal-01372037

Hermes, D.J. (1998). Measuring the Perceptual Similarity of Pitch Contours. *Journal of Speech, Language and Hearing Research, 41*, 73–82

Institut National du Cancer. (2019). Les cancers en France en 2018 – L'essentiel des faits et chiffres – Edition 2019. Consulté à l'adresse https://www.e-cancer.fr/Expertises-et-publications/Catalogue-des-publications/Les-cancers-en-France-en-2018-L-essentiel-des-faits-et-chiffres-edition-2019

Institut National du Cancer (2019). Synthèse – Estimations nationales de l'incidence et de la mortalité par cancer en France métropolitaine entre 1990 et 2018. Consulté à l'adresse https://www.e-cancer.fr/Expertises-et-publications/Catalogue-des-publications/Synthese-Estimations-nationales-de-l-incidence-et-de-la-mortalite-par-cancer-en-France-metropolitaine-entre-1990-et-2018

Kraaijenga, S.A.C., Oskam, I.M., Van Son, R.J.J.H., Hamming-Vrieze, O., Hilgers, F.J.M., Van den Brekel, M.W.M., Van der Molen, L. (2016). Assessment of voice, speech, and related quality of life in advanced head and neck cancer patients 10-years+ after chemoradiotherapy. *Oral Oncology*, *55*, 24-30. https://doi.org/10.1016/j.oraloncology.2016.02.001

Kreiman, J., Gerratt, B.R., Kempster, G.B., Erman, A., Berke, G.S. (1993). Perceptual evaluation of voice quality: review, tutorial, and a framework for future research. *Journal of Speech and Hearing Research, 36*(1), 21-40.

Kremer, J.M., Lederlé, E., Maeder, C. (2016). *Guide de l'orthophoniste* (Volume IV). Paris, France: Lavoisier-Médecine Sciences Publications.

La Ligue contre le Cancer, & Institut National du Cancer. (2018). *Les traitements des cancers des voies aérodigestives supérieures*.

Nocaudie, O., Astésano, C., Ghio, A., Lalain, M., Woisard, V. (2018). Evaluation de la compréhensibilité et conservation des fonctions prosodiques en perception de la parole de patients post-traitement de cancers de la cavité buccale et du pharynx. *Journées d'Etudes sur la Parole*. Aix-en-Provence, France. pp.196-204. https://hal.archives-ouvertes.fr/hal-01962272

Roux, G., Bertrand, R., Ghio, A., Astésano, C. (2016). Naïve listeners' perception of prominence and boundary in French spontaneous speech. *Speech Prosody*. https://hal.archives-ouvertes.fr/hal-01462259

Sicard, E., Mauclair, J., Woisard, V. (2017). *Etude de paramètres acoustiques des voix de patients traités pour un cancer ORL dans le cadre du projet C2SI*. Publication présentée aux 7èmes Journées de Phonétique Clinique, Paris, France. https://hal.archives- ouvertes.fr/hal-01510418/document

Schuster, M., Stelzle, F. (2012). Outcome measurements after oral cancer treatment: speech and speech-related aspects—an overview. *Oral and Maxillofacial Surgery*, *16*(3), 291-298. https://doi.org/10.1007/s10006-012-0340-y

Tribout, A. (2013). *Parler après une cordectomie : intelligibilité et qualité de vie* [Mémoire d'orthophonie]. Université Bordeaux Segalen.

Warren, R.M., Warren, R.P. (1970). Auditory illusions and confusions. Sci. Am.; 223, 30-36

Woisard, V., Espesser, R., Ghio, A., Duez, D. (2013). De l'intelligibilité à la compréhensibilité de la parole, quelles mesures en pratique clinique ? *Revue de Laryngologie Otologie Rhinologie*, *134*(1), 27-33. https://hal.archives-ouvertes.fr/hal-01486715

ANNEXES

Annexe 1 : Taux d'incidence et de mortalité par région anatomique dans les voies aérodigestives supérieures en 2018

Annexe 2 : Fiche consignes

Annexe 3 : Résultats du premier jury d'écoute

Annexe 4 : Classification TNM

Annexe 5 : Fiche évaluation avec consignes

Annexe 6 : Région anatomique des tumeurs en fonction du sexe

Annexe 7 : Traitements entrepris

Annexe 1 : Taux d'incidence et de mortalité par région anatomique dans les voies aérodigestives supérieures en 2018

TABLEAU 1 | Tumeurs solides : Cas incidents/décès estimés, taux d'incidence/de mortalité (TSM[1]) par localisation en 2018 et tendances évolutives (1990-2018 et 2010-2018) en France métropolitaine, chez l'homme

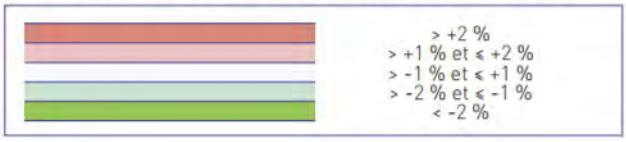
	Incidence						Mortalité					
	Situation en 2018		Variation annuelle moyenne (% [IC95])				Situation en 2018		Variation annuelle moyenne (% [IC95])			
Sites (et sous-types)	Nombre de nouveaux cas	Taux d'incidence[1]	1990-2018	IC95	2010-2018	IC95	Nombre de décès	Taux de mortalité[1]	1990-2018	IC95	2010-2018	IC95
Lèvre-bouche-pharynx[2]	10 055	18,3	-2,6	[-2,8 ; -2,5]	-1,9	[-2,4 ; -1,4]	2 898	4,9	-3,5	[-3,7 ; -3,4]	-2,8	[-3,2 ; -2,5]
Fosses nasales, sinus et oreilles[3]	552	1,0	-0,7	[-1,4 ; -0,1]	-0,7	[-2,2 ; 0,9]	-	-	-	-	-	-
Larynx	2 753	4,8	-3,1	[-3,4 ; -2,8]	-2,8	[-3,5 ; 2,1]	819	1,2	-6,3	[-6,5 ; -6,1]	-5,5	[-6,0 ; -4,9]

TABLEAU 2 | Tumeurs solides : Cas incidents/décès estimés, taux d'incidence/de mortalité (TSM[1]) par localisation en 2018 et tendances évolutives (1990-2018 et 2010-2018) en France métropolitaine, chez la femme

	Incidence						Mortalité					
	Situation en 2018		Variation annuelle moyenne (% [IC95])				Situation en 2018		Variation annuelle moyenne (% [IC95])			
Sites (et sous-types)	Nombre de nouveaux cas	Taux d'incidence[1]	1990-2018	IC95	2010-2018	IC95	Nombre de décès	Taux de mortalité[1]	1990-2018	IC95	2010-2018	IC95
Lèvre-bouche-pharynx[2]	3 637	5,8	1,8	[1,5 ; 2,1]	1,7	[0,9 ; 2,4]	924	1,2	-0,4	[-0,6 ; -0,2]	0,2	[-0,5 ; 0,8]
Fosses nasales, sinus et oreilles[3]	254	0,4	1,0	[0,1 ; 1,9]	1,0	[0,1 ; 1,9]	-	-	-	-	-	-
Larynx	407	0,7	0,0	[NC]	0,0	[NC]	131	0,2	-2,4	[-2,8 ; -1,9]	-2,3	[-3,4 ; -1,1]

IC95 : intervalle de confiance à 95 %
NC : Non calculé

Légende : Variation annuelle moyenne 1990-2018 ou 2010-2018



Sources : Incidence : Données des registres des cancers du réseau FRANCIM ; Mortalité : Données du CépiDc – Inserm. Pour chaque site, sous-site et sous-type, la liste de codes sélectionnés est présentée dans le chapitre matériel et méthode du rapport.
[1] TSM : Taux standardisés selon la structure d'âge de la population mondiale et exprimés pour 100 000 personnes-années.
[2] Site présentant des subdivisions par sous-sites topographiques/sous-types histologiques (avertissement : la somme des estimations des sous-sites/sous-types peut différer légèrement de celle du site – voir méthode).
[3] Nouveau site par rapport à la précédente étude 1980-2012 (Binder et al 2013).

(Institut National du Cancer, 2019)

Annexe 2 : Fiche consignes

Consigne : Vous allez écouter des enregistrements de voix. Il vous est demandé, après chaque écoute, de cocher la case marquant le degré de sévérité de l'intelligibilité perçu.

Enregistrement 1					
Grade d'intelligibilité	Sévère	Sévère-médium	Médium	Médium-léger	Léger

Enregistrement 2					
Grade d'intelligibilité	Sévère	Sévère-médium	Médium	Médium-léger	Léger

Enregistrement 3					
Grade d'intelligibilité	Sévère	Sévère-médium	Médium	Médium-léger	Léger

Enregistrement 4					
Grade d'intelligibilité	Sévère	Sévère-médium	Médium	Médium-léger	Léger

Enregistrement 5					
Grade d'intelligibilité	Sévère	Sévère-médium	Médium	Médium-léger	Léger

Enregistrement 6					
Grade d'intelligibilité	Sévère	Sévère-médium	Médium	Médium-léger	Léger

Enregistrement 7					
Grade d'intelligibilité	Sévère	Sévère-médium	Médium	Médium-léger	Léger

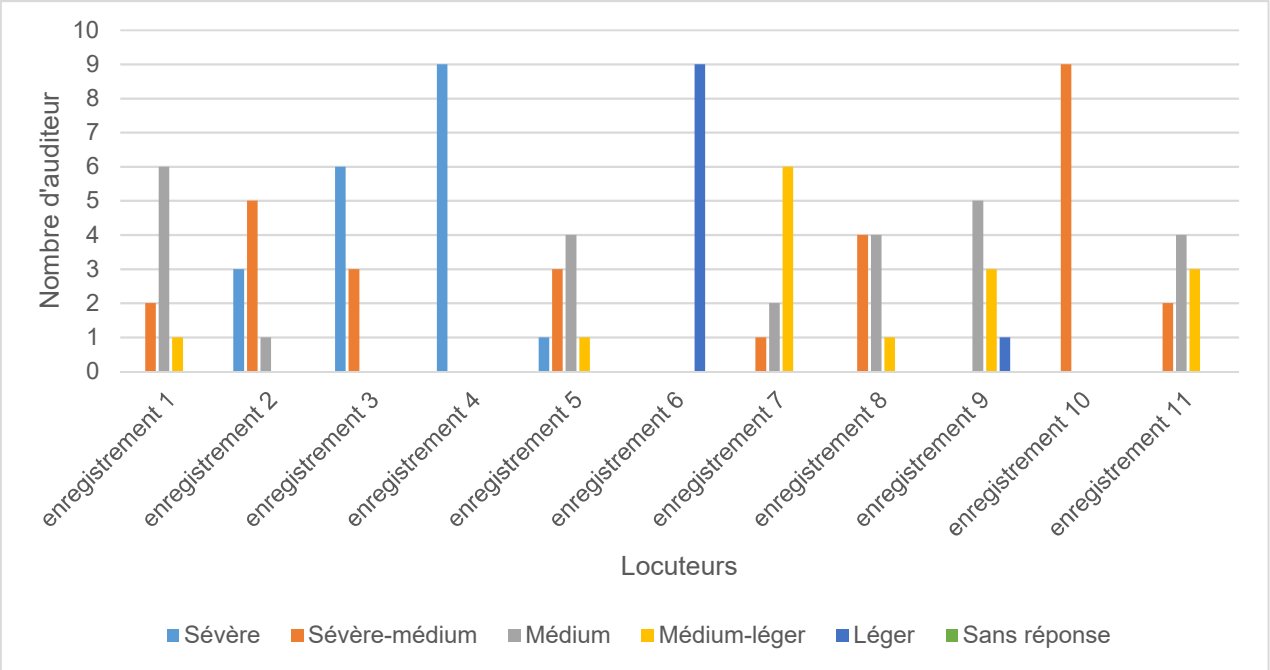
Enregistrement 8					
Grade d'intelligibilité	Sévère	Sévère-médium	Médium	Médium-léger	Léger

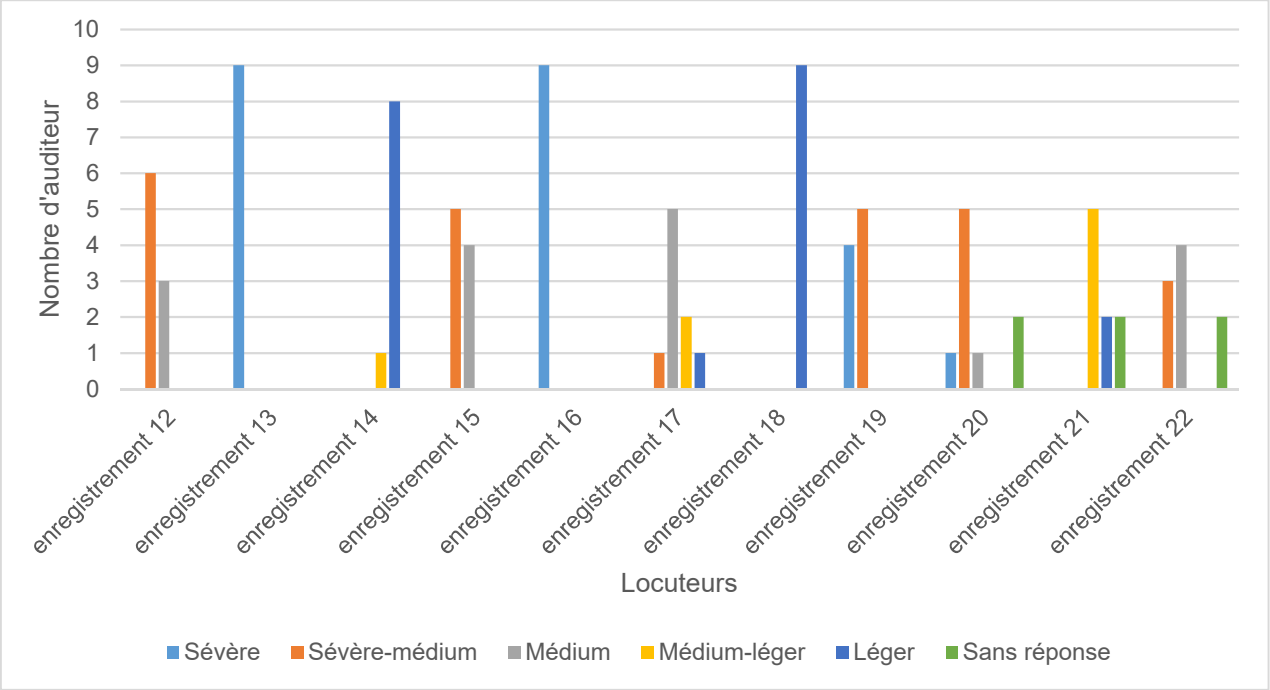
Enregistrement 9					
Grade d'intelligibilité	Sévère	Sévère-médium	Médium	Médium-léger	Léger

Enregistrement 10					
Grade d'intelligibilité	Sévère	Sévère-médium	Médium	Médium-léger	Léger

Enregistrement 11					
Grade d'intelligibilité	Sévère	Sévère-médium	Médium	Médium-léger	Léger

Annexe 3 : Résultats du premier jury d'écoute





Nous pouvons constater que pour les cas extrêmes (grave et léger), il y a une quasi-unanimité des auditeurs. En ce qui concerne les catégories intermédiaires, les avis sont plus partagés.

A noter pour les trois derniers enregistrements, le nombre d'auditeurs est incomplet, deux n'ayant pas pu terminer la tâche demandée.

Annexe 4 : Classification des tumeurs TNM

Cavité buccale	
T1	Tumeur inférieure ou égale à 2cm
T2	Tumeur supérieure à 2cm et inférieure à 4cm
T3	Tumeur supérieure à 4cm
T4	Tumeur s'étendant aux structures voisines (muscles, os)

Cavum	
T1	Tumeur limitée à une région du cavum de moins de 2 cm
T2	Tumeur envahissant deux régions du cavum
T3	Tumeur étendue à la région des fosses nasales et/ou de l'oropharynx
T4	Tumeur étendue à la base du crâne / nerfs crâniens

Oropharynx	
T1	Tumeur inférieure ou égale à 2cm
T2	Tumeur supérieure à 2cm et inférieure à 4cm
T3	Tumeur supérieure à 4cm
T4	Tumeur s'étendant aux structures voisines (muscles, os)

Hypopharynx	
T1	Tumeur limitée à une région de l'hypopharynx de moins de 2cm
T2	Tumeur envahissant plus d'une région de l'hypopharynx ou une région voisine sans fixation du larynx
T3	Tumeur envahissant plus d'une région de l'hypopharynx ou une région voisine avec fixation du larynx
T4	Tumeur étendue à l'os, au cou

Envahissement ganglionnaire	
N0	Absence de ganglions cliniquement métastatiques
N1	Ganglions homolatéraux mobiles
N2	Ganglions controlatéraux ou bilatéraux mobiles
N3	Ganglions fixés

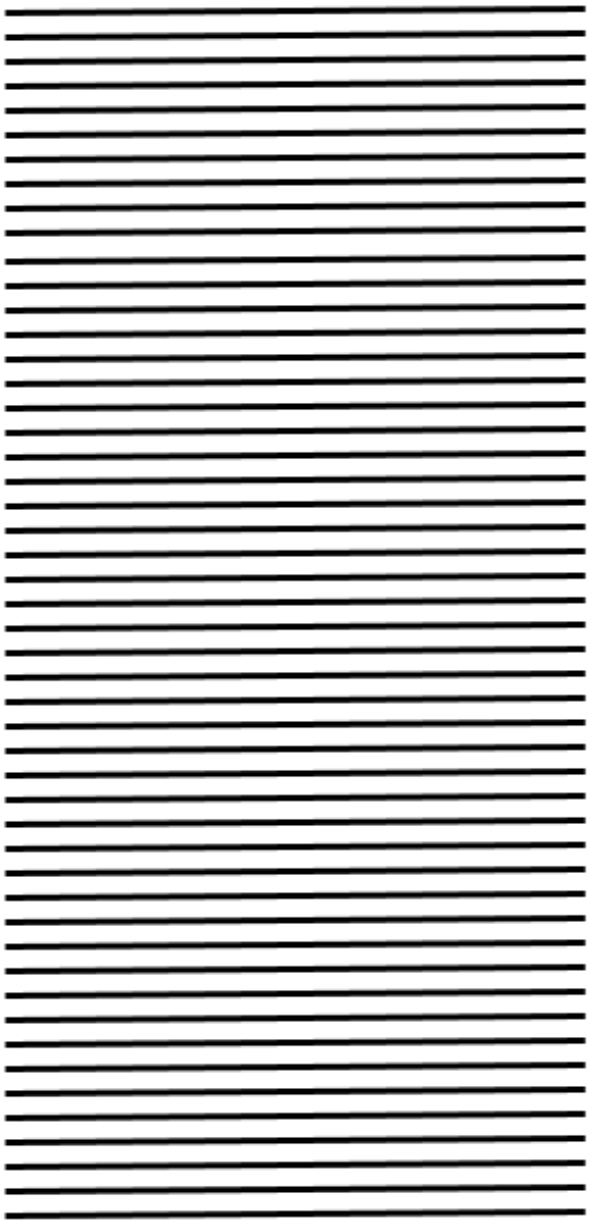
Métastases	
M0	Pas de signe de métastases à distance
M1	Présence de métastase(s) à distance

Annexe 5 : Fiche évaluation avec consignes

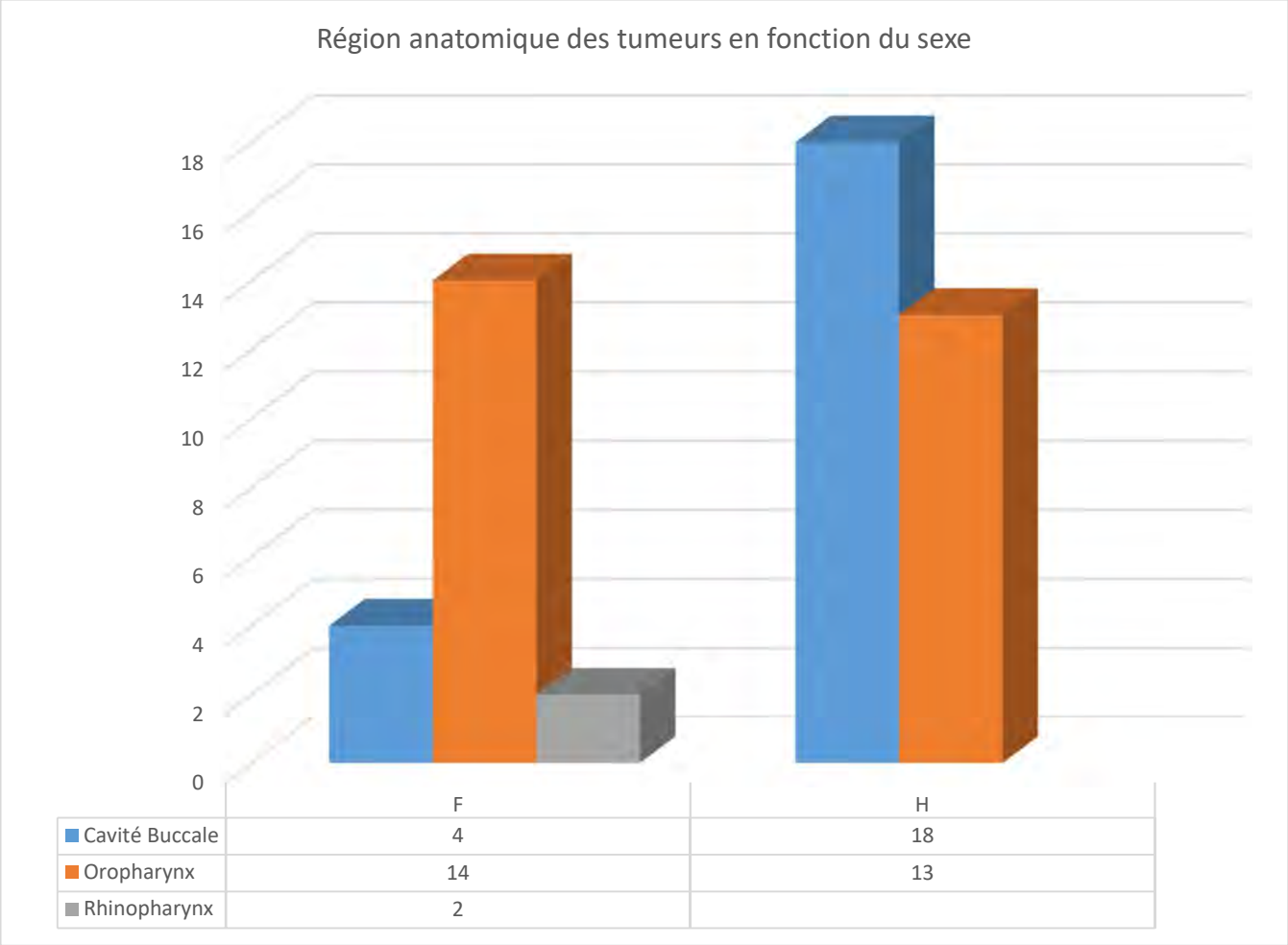
Consignes :

Vous allez écouter des enregistrements de voix.

Il vous est demandé après chaque écoute de placer sur une droite un trait marquant le degré d'altération globale du signal sonore, en sachant que l'extrémité gauche de l'échelle correspond à « aucune altération » et que l'extrémité droite correspond à une « altération très sévère ».



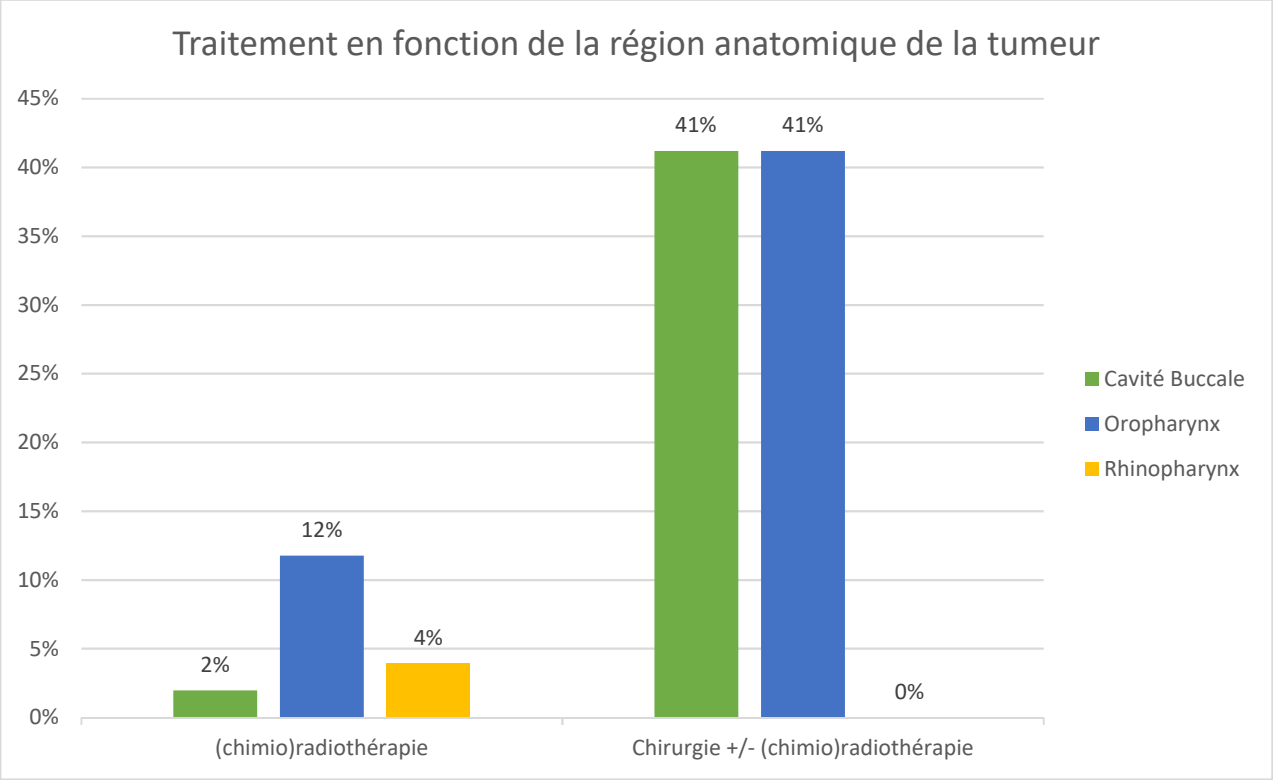
Annexe 6 : Région anatomique des tumeurs en fonction du sexe

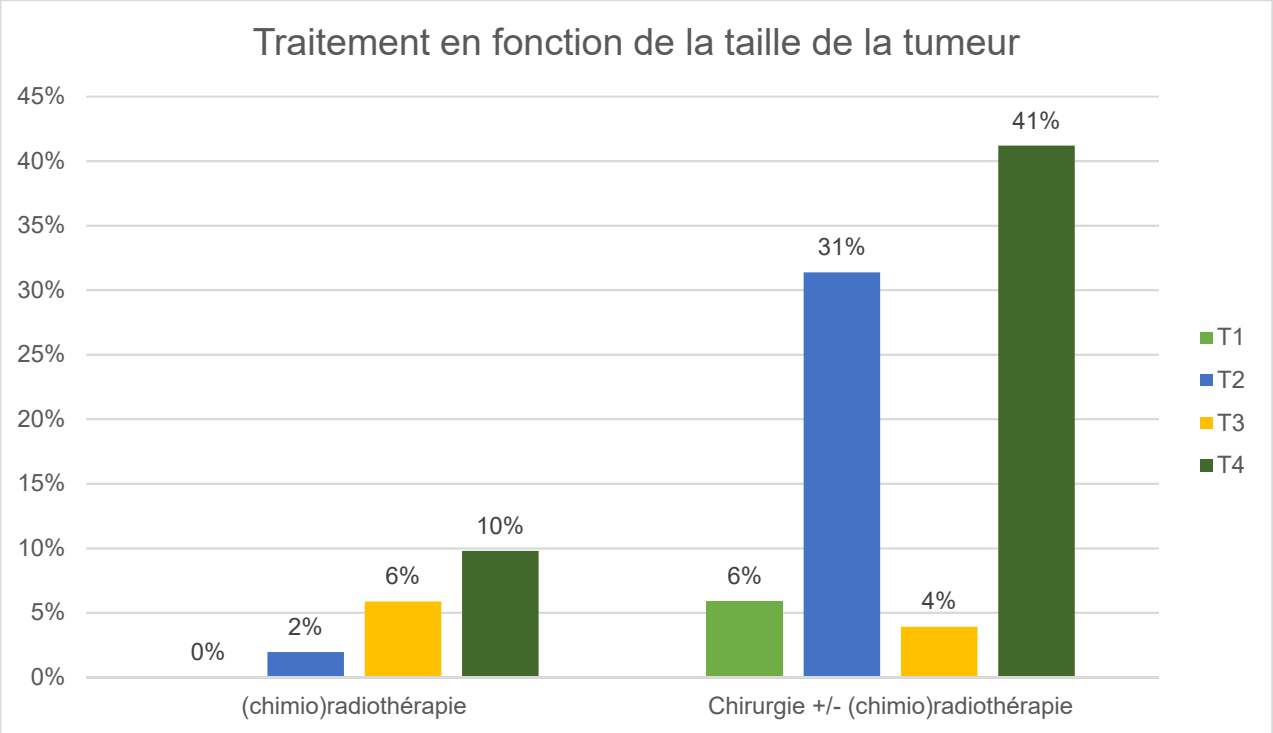


	F	H
Cavité Buccale	4	18
Oropharynx	14	13
Rhinopharynx	2	

Dans notre population d'étude, nous constatons que les tumeurs de l'oropharynx sont distribuées de façon équivalente entre hommes et femmes. Les tumeurs de la cavité buccale concernent majoritairement les hommes et les deux cas de cancers du rhinopharynx concernent les femmes.

Annexe 7 : Traitements entrepris





RESUME

Introduction : Grâce aux progrès dans le domaine de la recherche médicale, la mortalité liée aux cancers des VADS recule. L'allongement de la durée de vie rend prioritaire la prise en charge des séquelles dues aux traitements ou aux effets secondaires tardifs. L'évaluation des troubles de la parole est basée en clinique sur des évaluations perceptives malgré les biais et limites connus. Le développement d'analyse par traitement automatique vise à pallier ces limites.

Objectif : Ce travail vise à estimer si l'analyse automatique et l'analyse perceptive concourent à une évaluation convergente de la sévérité d'altération de la parole. De plus, ce travail vise également à estimer si cette évaluation pourrait être réalisée au travers d'un protocole simplifié en termes de tâches pour le patient.

Matériel et méthode : Notre étude porte sur des enregistrements rétrospectifs de 51 patients traités pour un cancer de la cavité buccale ou de l'oropharynx. De par leur production, les patients ont été répartis dans cinq groupes de sévérité de l'intelligibilité. Après sélection des tâches de lecture de texte (premier paragraphe du texte de « La chèvre de Monsieur Seguin ») et de phrases (les 10 phrases de la première liste de Combescure), ces enregistrements ont fait l'objet d'une évaluation perceptive et d'un traitement automatique. L'évaluation perceptive a été réalisée par un jury d'auditeurs naïfs devant noter le niveau d'altération global du signal sonore, sur une échelle de 0 (« aucune altération ») à 10 (« altération sévère »). L'analyse automatique a été réalisée en parallèle en utilisant un algorithme issu de la reconnaissance automatique de phonèmes. Les scores de vraisemblance obtenus n'ont pas de valeur seuil, contrairement aux scores perceptifs. Ces scores ont été analysés en termes de correspondance avec le degré de sévérité et les uns aux autres. Enfin, les facteurs de corrélation ont été calculés entre les scores obtenus à partir de l'énoncé de l'extrait de texte et chacune des phrases (ou combinaisons de phrases) afin de tester dans quelle mesure une réduction de la tâche de lecture des locuteurs est compatible avec une estimation suffisante de la qualité de parole produite.

Résultats : Les scores obtenus via un jury d'auditeurs naïfs sont variables. Malgré cette variabilité, l'ensemble des données conduit à des résultats cohérents en termes d'affectation aux différents groupes de sévérité avec néanmoins, une difficulté plus grande de noter des

locuteurs présentant des troubles de sévérité intermédiaire. Le facteur de corrélation calculé entre les scores perceptifs de l'extrait de texte et l'ensemble des phrases est de 0,85. Une estimation de ce facteur a également été calculée pour chaque phrase ainsi que différentes combinaisons (entre 0,64 et 0,85). L'analyse automatique réalisée en parallèle permet d'individualiser de manière significative le niveau de sévérité « légère » aux autres niveaux. De plus, on obtient une corrélation entre les scores automatique et perceptif de -0,64 pour le texte et -0,67 pour l'ensemble des phrases, facteurs pouvant être de meilleure qualité si les valeurs aberrantes sont écartées (-0,75 et -0,81 respectivement). De même que pour l'analyse perceptive, une estimation du facteur de corrélation a été calculée pour chaque phrase ainsi que différentes combinaisons.

Conclusion : Au vu des résultats, il serait possible de réduire la tâche de lecture des locuteurs pour évaluer leur sévérité d'altération de la parole via une analyse automatique et perceptive. Les résultats obtenus par analyse automatique montrent : i) une complémentarité des résultats obtenus avec l'analyse perceptive. ii) l'analyse automatique semble sensible à la qualité des enregistrements. De plus, l'ensemble des résultats indiquent que : i) le premier paragraphe de « La chèvre de Monsieur Seguin » peut être substitué par l'utilisation de phrases de Combescure. ii) la tâche d'évaluation pourrait être réduite à l'aide d'une combinatoire de phrases comprenant entre 25-30 mots. Il reste à confirmer ces résultats et à optimiser la composition des phrases à soumettre au locuteur en termes de nombre de mots et/ou de syllabes.