Nouveaux biomarqueurs vocaux pour la détection automatique de la somnolence

Vincent Martin

THÈSE

présentée à

L'UNIVERSITÉ DE BORDEAUX

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

par

Vincent MARTIN

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : INFORMATIQUE

Nouveaux biomarqueurs vocaux pour la détection automatique de la somnolence

---

Date de soutenance : Mercredi 8 juin 2022

Devant la commission d'examen composée de :

Pr. Corinne	Fredouille	PU - Univ. d'Avignon	Rapportrice[†]
Pr. Isabel	Trancoso	PU - Univ. de Lisbonne	Rapportrice
Dr. Pierre-Alexis	Geoffroy	MCU-PH (HDR) - Univ. de Paris	Rapporteur
Dr. Véronique	Delvaux	CQ FNRS - Univ. de Mons	Examinatrice
Dr. Guy	Fagherazzi	Dir. du Dpt. de Médecine de Précision (HDR) Luxembourg Institute of Health	Examinateur
Dr. Jean-Luc	Rouas	CR CNRS (HDR) - Univ. de Bordeaux	Directeur
Pr. Pierre	Philip	PU-PH - Univ. de Bordeaux	Co-directeur
Dr. Jean-Arthur	Micoulaud-Franchi	MCU-PH (HDR) - Univ. de Bordeaux	Encadrant, Invité

[†] Présidente du jury

– 2022 –

Titre

Nouveaux biomarqueurs vocaux pour la détection automatique de la somnolence

Résumé

La voix est un des outils les plus prometteurs de la médecine numérique. En association avec les compagnons virtuels médicaux, l'estimation de symptômes à partir de marqueurs vocaux permettra à la fois le suivi à domicile de patients souffrant de maladies neuropsychiatriques chroniques, et l'accès à des conseils personnalisés d'hygiène de vie pour la population générale. La somnolence, présente dans de nombreuses pathologies et présentant une très forte prévalence à la fois chez les patients souffrant de maladies chroniques et en population générale, est un symptôme privilégié pour cette approche. L'objectif des travaux présentés dans ce manuscrit est ainsi de compléter les informations collectées par les assistants virtuels lors de l'interaction des sujets avec ceux-ci, en utilisant des marqueurs vocaux validés comme étant des marqueurs fiables de la somnolence. La démarche suivie est la suivante.

Dans un premier temps, nous introduisons les mécanismes de production de la voix et l'ensemble des pathologies qui peuvent interférer avec les différentes fonctions musculaires et neuro-musculaires impliquées, avec une attention particulière portée sur les méthodologies employées pour l'enregistrement et l'annotation des corpus utilisés.

Ensuite, nous tentons d'établir une définition consensuelle de la somnolence en utilisant trois dictionnaires de référence de la langue française ; deux approches de fouille de texte ; et enfin par l'intermédiaire d'une revue générale des outils conçus pour la mesurer.

Nous présentons ensuite notre propre corpus de patients atteints d'hypersomnies, enregistrés au pôle universitaire de médecine du sommeil du CHU de Bordeaux sur une tâche de lecture à voix haute, annotés avec des mesures de somnolence à la fois subjectives (questionnaires) et objectives (latence d'endormissement au Test Itératif de Latence d'Endormissement – TILE) validées par les médecins du CHU. Ce corpus est ensuite comparé avec les autres corpus de l'état de l'art sur la détection de la somnolence dans la voix, à partir desquels nous proposons des recommandations sur l'élaboration de tels corpus. Puis, à l'aide d'une étude perceptuelle, nous validons l'utilisation de la base TILE pour la détection de la somnolence dans la voix.

Sur la base de ce corpus, nous élaborons ensuite quatre catégories de descripteurs vocaux, mesurant deux dimensions de l'impact de la somnolence sur la voix et la production de parole. D'une part, nous étudions des marqueurs de qualité acoustique de la voix ; d'autre part nous concevons des marqueurs de qualité de lecture, divisés en trois sous-catégories : les erreurs de lecture faites par les patients, leur automatisation à travers les erreurs faites par des systèmes de reconnaissance automatique de la parole, et enfin les durées et emplacements des pauses de lecture. Ces marqueurs sont validés sur différentes formes de somnolence (objective et subjective).

Enfin, nous proposons une méthodologie pour entraîner un classifieur dans la visée d'une utilisation clinique de ces descripteurs vocaux pour la détection de trois symptômes liés à la somnolence excessive. Nous proposons une analyse détaillée des résultats obtenus et des descripteurs employés par le classifieur. Pour aller plus loin, nous proposons ensuite de rapprocher le problème de classification de la réalité du raisonnement clinique en classifiant deux syndromes dérivés des précédents symptômes. Enfin, dans cette même direction, nous proposons des perspectives de recherche autour des réseaux de symptômes, dans la cadre de la recherche en médecine numérique sur la somnolence et sur la psychiatrie numérique de manière plus générale.

Mots-clés

Somnolence ; Descripteurs vocaux ; Qualité acoustique de la voix ; Système de transcription automatique de la parole ; Pauses de lecture ; Médecine numérique.

Title

New speech biomarkers for automatic sleepiness detection

Abstract

Voice is one of the most promising tools in digital medicine. In association with virtual medical companions, the estimation of symptoms based on voice features will allow both home monitoring of patients suffering from chronic neuropsychiatric diseases and access to personalized lifestyle advice for the general population. Sleepiness, occuring in many pathologies and being very prevalent both in patients suffering from chronic diseases and in the general population, is a key symptom for this approach. The objective of the work presented in this manuscript is thus to complete the information collected by virtual assistants during the interaction of the subjects with them, by using vocal markers validated as being reliable markers of sleepiness. Our approach is the following.

First, we introduce the mechanisms of voice production and the different pathologies that can interfere with the involved muscular and neuro-muscular functions, with a focus on the methodologies used for the recording and annotation of the corpora.

Then, we attempt to establish a consensual definition of sleepiness using three reference dictionaries of the French language; two text mining approaches; and finally through an umbrella review of tools designed to measure it.

Subsequently, we present our own corpus of patients with hypersomnia, recorded at the sleep medicine center of the Bordeaux University Hospital on a reading aloud task, annotated with both subjective (questionnaires) and objective (sleep latency to the Multiple Sleep Latency Test) measures of sleepiness validated by the physicians of the University Hospital. This corpus is then compared with other state-of-the-art corpora on voice sleepiness detection, from which we propose recommendations on the development of such corpora. Then, using a perceptual study, we validate the use of the MSLT database for the detection of sleepiness in speech.

Based on this corpus, we develop four categories of speech features, measuring two dimensions of the impact of sleepiness on speech. On the one hand, we study markers of acoustic voice quality; on the other hand, we design markers of reading quality, divided into three subcategories: reading errors made by patients, their automation through errors made by automatic speech recognition systems, and finally the durations and locations of reading pauses. These features are validated on different forms of sleepiness (objective and subjective).

Finally, we present a methodology to train a classifier for the clinical use of these speech features for the detection of three symptoms related to sleepiness. We carry out a detailed analysis of the obtained results and of the descriptors used by the classifier. To go further, we then propose to bring the classification problem closer to the reality of clinical reasoning by classifying two syndromes derived from the previous symptoms. Finally, in this same direction, we consider research perspectives around symptom networks, in the framework of digital medicine research on sleepiness and, in a more general way, on digital psychiatry.

Keywords

Sleepiness; Vocal features; Acoustic quality of voice; Automatic speech recognition; Reading pauses; Digital medicine.

Remerciements

À mes rapportrices et rapporteur, Mmes Corinnne Fredouille et Isabel Trancoso et M. Pierre-Alexis Geoffroy, pour avoir accepté la responsabilité de lire et de rapporter cette thèse, ainsi que Mme Véronique Delvaux et M. Guy Fagherazzi pour leur rôle respectif d'examinatrice et d'examinateur de ces travaux.

À MM. Jean-Philippe Dominger et Xavier Blanc, directeurs successifs du LaBRI, ainsi que MM. Andreas Hartman et Guillaume Blin, directeurs successifs de l'école doctorale en mathématiques et informatiques, pour le cadre de travail et d'apprentissage qu'ils m'ont offerts pour la réalisation de cette thèse. Je remercie à ce titre tous les personnels administratifs et d'appui à la recherche à la fois au LaBRI et du SANPSY, ainsi que les infirmières et infirmier du service Universitaire de médecine du Sommeil de Bordeaux, qui m'ont réservé un accueil chaleureux au service de médecine du sommeil lors de la collecte de la base TILE.

À mon directeur de thèse, Jean-Luc, pour la relecture précise et impitoyable de ce document, mais surtout pour ton soutien indéfectible (même dans mes idées les plus farfelues) et tes conseils avisés, année après année.

À mon co-directeur de thèse, le Pr. Pierre Philip, pour son exigence, sa rigueur scientifique et son accompagnement.

À Jean-Arthur, pour ta disponibilité et ta grande bienveillance à mon égard, à la fois dans le cadre de nos travaux sur la somnolence que lors du DIU de philosophie de la psychiatrie.

À Régis, pour ta disponibilité et ta didactique sur le concept de somnolence, et pour la co-rédaction de l'article associé, durant laquelle j'ai beaucoup appris.

À Christophe, pour ton optimisme débordant et nos discussions autour de l'apprentissage automatique, de l'épistémologie, de la psychiatrie et des réseaux de symptômes.

À mes collègues et anciens collègues de bureau : Reda, Huy-Dung, Florian, Pierre (sans qui les figures de ce manuscrit seraient moins nombreuses, et encore en nuances de gris), les membres successifs de l'AFODIB, mais aussi tous les collègues du LaBRI (et plus spécialement de l'équipe I&S et de TAD) et du SANPSY pour leurs échanges, riches et formateurs. De même, je remercie tous les collègues aux côtés desquels j'ai pu enseigner, à l'Université comme à l'IUT de Bordeaux.

Aux nombreux stagiaires de divers horizons que j'ai la chance de pouvoir encadrer, avec lesquels j'ai appris énormément : Pierre Thivel, Gabrielle Chapouthier, Mathilde Rieant, Agathe Basse, Benoît Caudron, Marie Huillet, Aymeric Ferron, Brice Arnaud.

Aux acteurs parfois oubliés mais essentiels à l'aboutissement de ces nombreuses années d'études que sont mes anciens professeurs de lycées (Mmes Chermette et Voirin, MM. Jay, Provost, ...), de classes préparatoires (Mmes Camus, Ponchart, MM. Bordes, Casseau, Stoki, Lhermitte, ...) ou encore d'école d'ingénieur (notamment MM. Simond et Reynal), à qui je dois mes méthodes de travail et mon appétence des sciences. De même, je souhaite ici remercier mes encadrants de stages de recherche, qui ont su créer et confirmer mon adéquation avec le milieu académique : MM. et Mmes Yannick Jeantet, Sylvain Reynal, Yoon Cho et Hélène Papadopoulos.

Aux amis qui m'ont soutenu pendant ces trois années : Simon, Maxime, Victor, Réhane, Hugo, Quentin, Stan, ... mais aussi tous les musiciens et danseurs rencontrés en bals ou en sessions, soupapes indispensables à un travail scientifique intense.

À ma famille, et particulièrement à mes parents, pour qui l'éducation a toujours été une priorité. Cette thèse est aussi le résultat de vos efforts.

À Seb, pour ton soutien sans faille dans tous les instants.

Sommaire

Introduction générale

Smartphones et interaction vocale

Alors qu'il aura fallu 46 ans à l'électricité, installée aux États-Unis en 1873, pour être adoptée par plus de 25% de la population et 7 ans entre l'inauguration d'internet en 1991 et son utilisation par un étatsunien sur quatre, l'adoption des smartphones (2007) et des assistants vocaux (2016) par un quart de la population américaine actuelle s'est faite en moins de deux ans (Topol, 2019, p. 139). En 2022, le nombre d'utilisateurs de smartphones est estimé à 6.2 milliards, possédés par 80% de la population mondiale [1]. Il est notamment possible d'interagir avec ceux-ci par la voix, complétant le panel de dispositifs intégrant des technologies vocales qui prend une place grandissante dans nos quotidiens. Par exemple, il est estimé qu'en 2018, 71% de la génération « millenials » utilisent les technologies basées sur la voix de façon quotidienne (Sorensen, 2019). Parler est en effet plus rapide et conduit à moins d'erreurs de contenu que de taper la même information sur un clavier, aussi bien dans des langues syllabiques (comme par ex. le français) que basées sur des idéogrammes (chinois) (Ruan *et coll.*, 2018). De plus, cette augmentation de la vitesse de codage de l'information par la voix, observée à la fois sur des tâches de production et de transcription, est accompagnée d'une diminution de la fatigue physique (telle que mesurée par l'index de charge de tâche de la NASA-TLX) (Foley *et coll.*, 2020).

La voix comme support de la santé numérique

L'utilisation de la voix pour interagir avec la technologie étant de plus en plus adoptée au quotidien par la population générale, la conception et l'implémentation de biomarqueurs vocaux pour le diagnostic ou le suivi de maladies ou de symptômes est une des applications les plus prometteuses de la santé numérique (Fagherazzi, 2021; The IQVIA Institute for Human Data Science, 2021). En effet, contrairement à d'autres moyens de mesures plus traditionnels comme la vidéo ou l'électroencéphalographie, l'enregistrement de signaux vocaux est peu cher (à la fois en termes financiers et humains), ne nécessite pas de ressources computationnelles importantes (permettant d'envisager un traitement en temps réel), et a une bonne robustesse aux environnements bruités. Ces avantages en font le candidat idéal pour être implémenté en dehors des laboratoires de recherches, aussi près que possible des patients, en conditions écologiques – comme par exemple dans une application smartphone de compagnon virtuel.

Smartphones et assistants médicaux virtuels

L'utilisation massive des smartphones par la population a conduit au développement de nombreux compagnons médicaux virtuels, permettant le suivi de pathologies aussi variées que la dépression (Philip *et coll.*, 2017), le stress post-traumatique (Lucas *et coll.*, 2017), l'apnée du sommeil (Dupuy *et coll.*, 2021), le cancer du sein (Beveridge et Fox, 2006), le diabète (Griol *et coll.*, 2013) ou encore l'hypertension (Giorgino *et coll.*, 2005). Une revue de ces différents dispositifs est proposée par Laranjo *et coll.* (2018).

Ces applications permettent à la fois de déplacer le lieu de suivi des patients des hôpitaux et cliniques vers leur domicile, proposant un suivi régulier au lieu des rendez-vous ponctuels et épisodiques proposés dans les services spécialisés, et de récolter les symptômes des patients dans des conditions au plus proches de leur vie quotidienne, par opposition aux biais induits par le milieu hospitalier. De plus, pour les troubles psychiatriques ou l'addiction (Auriacombe *et coll.*, 2021), ces applications permettent aux patients de s'évaluer sans risque de jugement

1. https://www.statista.com/statistics/330695/number-of-smartphone-users-worldwide/

du clinicien, favorisant le processus d'*empowerment* de ces patients (Guelfi *et coll.*, 2021, p.797). Par ailleurs, elles permettent la prodigation de conseils personnalisés à partir des informations renseignées par les patients, et la création d'un profil numérique du patient, que le clinicien peut éventuellement récupérer lors des consultations.

C'est par exemple ce qui est proposé par l'application gratuite KANOPÉE, qui permet le suivi de l'insomnie, de la fatigue, de l'anxiété et de l'addiction (alcool, cigarette, cannabis) au quotidien (Dupuy *et coll.*, 2021; Philip *et coll.*, 2020). Après une première phase d'estimation de gravité des symptômes, l'application propose de remplir un agenda du sommeil (problèmes liés au sommeil) ou de consommations (addictions), et prodigue des conseils sur les comportements à adopter face à ces plaintes. Si au bout de la période d'utilisation de l'application, les symptômes persistent malgré le suivi des conseils prodigués par l'application, l'utilisateur est invité à prendre contact avec un médecin spécialisé.

Parmi les nombreux symptômes qu'il est possible de suivre avec ces applications, la somnolence est un symptôme commun à de nombreuses pathologies (Jike *et coll.*, 2018; Scott *et coll.*, 2021). Elle est donc une candidate idéale pour l'implémentation de ces technologies.

La somnolence : un problème de santé publique

La somnolence est un état psychophysiologique normal dont la majorité des individus font l'expérience sur une période de 24 heures (Shen *et coll.*, 2006; Ohayon, 2008). Cependant, lorsqu'elle survient à des moments inappropriés, avec une fréquence élevée ou avec des conséquences importantes, la somnolence est généralement considérée comme un trouble appelé *somnolence excessive* (Ohayon *et coll.*, 2012). La somnolence excessive est un symptôme potentiellement dangereux pour la santé individuelle et publique qui fait l'objet d'une attention scientifique, sociale et politique croissante, notamment en ce qui concerne l'accidentologie routière (Bioulac *et coll.*, 2017) : par son impact sur nos capacités d'attention et de vigilance, la somnolence est responsable d'un accident mortel sur trois sur autoroute en France, et multiplie par huit le risque d'accident [2].

De plus, la somnolence excessive est souvent associée à un large éventail de maladies, notamment les troubles du sommeil, métaboliques, cardiovasculaires, neurologiques et psychiatriques, entraînant un handicap et un risque accru de mortalité (Jike *et coll.*, 2018; Scott *et coll.*, 2021). La somnolence excessive est également couramment associée à des répercussions sociales et économiques, constituant ainsi un problème de santé publique important (Barnes et Watson, 2019; Léger et Stepnowsky, 2020).

La somnolence excessive peut refléter différentes pathologies, dont une hygiène ou des comportements inadéquats de sommeil, les effets sédatifs d'une substance, mais aussi divers troubles sous-jacents, en particulier des troubles du sommeil. La somnolence excessive est ainsi l'une des plaintes les plus courantes et les plus centrales des personnes préoccupées par leur sommeil (Gauld *et coll.*, 2021).

Jusqu'à présent, aucun consensus clair n'a été établi quant à la définition précise du concept de *somnolence* et aux seuils à partir desquels la somnolence est considérée comme *excessive*. Cela explique, en partie, les grandes différences dans l'estimation de sa prévalence. En effet, rien que dans la revue réalisée par Young (2004), la prévalence de la somnolence excessive varie de 2,5% (somnolence excessive définie comme « causant des problèmes au travail », 1186 adultes polonais) à plus de 40% (somnolence excessive définie comme « fatigue quotidienne associée à au moins un problème diurne », 3328 adultes américains). Sur 15929 citoyens américains, Ohayon *et coll.* (2012) ont trouvé une prévalence allant de 4,7% (somnolence

2. https://www.securite-routiere.gouv.fr/dangers-de-la-route/la-fatigue-et-la-conduite

excessive rapportée « au moins trois fois par semaine pendant au moins trois mois ») à 27,8% (somnolence excessive rapportée par les sujets), tandis que Jaussent *et coll.* (2017) rapportent une prévalence de la somnolence excessive de 32,9% (score à l'échelle de somnolence d'Epworth > 10) et une prévalence de la persistance de la somnolence excessive chez 32,6% des sujets, sur la base d'un échantillon de 2167 sujets exempts de troubles d'hypersomnie centrale et travailleurs de jour.

Objectif et méthode

Élaborer un système de détection de la somnolence dans la voix bénéficiera à la fois aux domaines industriels orientés sur des tâches nécessitant des niveaux d'attention élevés (par ex. la défense, le nucléaire ...), mais aussi à la future médecine numérique, qui pourra s'appuyer sur cet outil pour détecter un symptôme majeur de nombreuses pathologies.

L'objectif des travaux présentés dans ce manuscrit est de compléter les informations collectées par les assistants virtuels lors de l'interaction des sujets avec eux, en utilisant des marqueurs vocaux validés comme étant des marqueurs fiables de la somnolence.

Pour cela, nous nous proposons dans ce manuscrit une première étape vers cet objectif : la validation de marqueurs vocaux, en conditions cliniques et contrôlées, sur les niveaux de somnolence de patients atteints de maladies chroniques du sommeil, qui est une population cible pour le suivi à domicile grâce à des agents conversationnels.

Dans cette visée, nous proposons dans la partie I une explication succincte des mécanismes moteurs et neuromoteurs impliqués dans la production vocale, et une revue de la littérature des pathologies étudiées par la communauté du traitement du signal vocal, et plus particulièrement des méthodologies de conception des bases de données utilisées.

Nous proposons dans la partie II d'établir une définition consensuelle de la *somnolence*, à partir de définitions de dictionnaires de langue courante, d'approches de fouilles de texte et par le biais d'une revue générale des outils conçus pour la mesurer. À partir des résultats des approches précédentes et de la littérature spécialisée sur le sujet, nous établissons ensuite un schéma relationnel des différentes construits adjacents à la somnolence.

La partie III introduit le corpus que nous avons enregistré au pôle universitaire de médecine du sommeil du CHU de Bordeaux, que nous comparons ensuite aux autres bases de données existant dans le champ de la détection de la somnolence dans la voix. Nous proposons, à partir de ces comparaisons, un ensemble de recommandations sur la construction de tels corpus. Enfin, nous concluons cette partie par une étude perceptuelle de la somnolence dans la voix, qui valide l'utilisation de cette base de données pour la détection automatique de la somnolence et propose de premiers résultats sur les caractéristiques des annotateurs et des locuteurs qui peuvent influencer la perception de la somnolence à partir d'échantillons vocaux.

Une fois le corpus collecté, nous proposons dans la partie IV quatre familles de descripteurs vocaux, dont nous validons la pertinence pour la détection de la somnolence : des descripteurs acoustiques, les erreurs de lecture, les erreurs faites par des systèmes de transcription automatique, et les durées et emplacements des pauses de lecture. Ce travail étant fait en collaboration avec des médecins, l'interprétabilité des marqueurs est impérative et impose des contraintes fortes sur les marqueurs conçus.

Enfin, nous finissons par proposer dans la partie V une méthodologie pour entraîner un classifieur dans la visée d'une utilisation clinique, qui, à partir des descripteurs précédemment présentés, permet la détection de trois symptômes liés à la somnolence excessive. Nous proposons une analyse complète des résultats obtenus et des marqueurs utilisés par chaque

classifieur, dans une démarche explicative nécessaire au travail interdisciplinaire avec les médecins. Nous proposons ensuite d'aller plus loin en rapprochant la tâche de classification de la réalité du raisonnement clinique, en classifiant deux syndromes liés – eux aussi – à la somnolence excessive. Enfin, nous concluons dans cette lignée en proposant des perspectives de recherche autour des réseaux de symptômes comme nouveau paradigme pour la classification de pathologies, et plus spécifiquement dans le cadre de la psychiatrie numérique.

Bibliographie

Auriacombe, M., Fournet, L., Dupuy, L., Micoulaud-Franchi, J.-A., de Sevin, E., Moriceau, S., Baillet, E., Alexandre, J.-M., Serre, F., et Philip, P. (**2021**). "Effectiveness and Acceptance of a Smartphone-Based Virtual Agent Screening for Alcohol and Tobacco Problems and Associated Risk Factors During COVID-19 Pandemic in the General Population," Frontiers in Psychiatry **12**, 693687, doi: 10.3389/fpsyt.2021.693687.

Barnes, C. M., et Watson, N. F. (**2019**). "Why healthy sleep is good for business," Sleep Medicine Reviews **47**, 112–118, doi: 10.1016/j.smrv.2019.07.005.

Beveridge, M., et Fox, J. (**2006**). "Automatic generation of spoken dialogue from medical plans and ontologies," Journal of Biomedical Informatics **39**(5), 482–499, doi: 10.1016/j.jbi.2005.12.008.

Bioulac, S., Micoulaud-Franchi, J.-A., Arnaud, M., Sagaspe, P., Moore, N., Salvo, F., et Philip, P. (**2017**). "Risk of Motor Vehicle Accidents Related to Sleepiness at the Wheel : A Systematic Review and Meta-Analysis," Sleep **40**(10), doi: 10.1093/sleep/zsx134.

Dupuy, L., Micoulaud-Franchi, J.-A., et Philip, P. (**2021**). "Acceptance of virtual agents in a homecare context : Evaluation of excessive daytime sleepiness in apneic patients during interventions by continuous positive airway pressure (CPAP) providers," Journal of Sleep Research **30**(2), e13094, doi: 10.1111/jsr.13094.

Fagherazzi, G. (**2021**). "Do I sound sick ?," The Lancet Digital Health **3**(9), e534, doi: 10.1016/S2589-7500(21)00182-5.

Foley, M., Casiez, G., et Vogel, D. (**2020**). "Comparing Smartphone Speech Recognition and Touchscreen Typing for Composition and Transcription," dans *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–11, doi: 10.1145/3313831.3376861.

Gauld, C., Lopez, R., Morin, C., Geoffroy, P. A., Maquet, J., Desvergnes, P., McGonigal, A., Dauvilliers, Y., Philip, P., Dumas, G., et Micoulaud-Franchi, J.-A. (**2021**). "Symptom network analysis of the sleep disorders diagnostic criteria based on the clinical text of the ICSD-3," Journal of Sleep Research **0**(0), e13435, doi: 10.1111/jsr.13435.

Giorgino, T., Azzini, I., Rognoni, C., Quaglini, S., Stefanelli, M., Gretter, R., et Falavigna, D. (**2005**). "Automated spoken dialogue system for hypertensive patient home management," International Journal of Medical Informatics **74**(2-4), 159–167, doi: 10.1016/j.ijmedinf.2004.04.026.

Griol, D., Carbó, J., et Molina, J. M. (**2013**). "An Automatic Dialog Simulation Technique To Develop And Evaluate Interactive Conversational Agents," Applied Artificial Intelligence **27**(9), 759–780, doi: 10.1080/08839514.2013.835230.

Guelfi, J.-D., Rouillon, F., et Mallet, L. (**2021**). *Manuel de psychiatrie* (Elsevier Health Sciences).

Jaussent, I., Morin, C. M., Ivers, H., et Dauvilliers, Y. (**2017**). "Incidence, worsening and risk factors of daytime sleepiness in a population-based 5-year longitudinal study," Scientific Reports **7**(1), 1372, doi: 10.1038/s41598-017-01547-0.

Jike, M., Itani, O., Watanabe, N., Buysse, D. J., et Kaneita, Y. (**2018**). "Long sleep duration and health outcomes : A systematic review, meta-analysis and meta-regression," Sleep Medicine Reviews **39**, 25–36, doi: 10.1016/j.smrv.2017.06.011.

Laranjo, L., Dunn, A. G., Tong, H. L., Kocaballi, A. B., Chen, J., Bashir, R., Surian, D., Gallego, B., Magrabi, F., Lau, A. Y. S., et Coiera, E. (**2018**). "Conversational agents in healthcare : a systematic review," Journal of the American Medical Informatics Association **25**(9), 1248–1258, doi: 10.1093/jamia/ocy072.

Léger, D., et Stepnowsky, C. (**2020**). "The economic and societal burden of excessive daytime sleepiness in patients with obstructive sleep apnea," Sleep Medicine Reviews **51**, 101275, doi: 10.1016/j.smrv.2020.101275.

Lucas, G. M., Rizzo, A., Gratch, J., Scherer, S., Stratou, G., Boberg, J., et Morency, L.-P. (**2017**). "Reporting Mental Health Symptoms : Breaking Down Barriers to Care with Virtual Human Interviewers," Frontiers in Robotics and AI **4**, 51, doi: 10.3389/frobt.2017.00051.

Ohayon, M. M. (**2008**). "From wakefulness to excessive sleepiness : What we know and still need to know," Sleep Medicine Reviews **12**(2), 129–141, doi: 10.1016/j.smrv.2008.01.001.

Ohayon, M. M., Dauvilliers, Y., et Reynolds, C. F. (**2012**). "Operational definitions and algorithms for excessive sleepiness in the general population : implications for DSM-5 nosology," Archives of general psychiatry **69**(1), 71–79, doi: 10.1001/archgenpsychiatry.2011.1240.

Philip, P., Dupuy, L., Morin, C. M., de Sevin, E., Bioulac, S., Taillard, J., Serre, F., Auriacombe, M., et Micoulaud-Franchi, J.-A. (**2020**). "Smartphone-Based Virtual Agents to Help Individuals With Sleep Concerns During COVID-19 Confinement : Feasibility Study," Journal of Medical Internet Research **22**(12), e24268, doi: 10.2196/24268.

Philip, P., Micoulaud-Franchi, J.-A., Sagaspe, P., De Sevin, E., Olive, J., Bioulac, S., et Sauteraud, A. (**2017**). "Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders," Scientific Reports **7**(1), 426–456, doi: 10.1038/srep42656.

Ruan, S., Wobbrock, J. O., Liou, K., Ng, A., et Landay, J. A. (**2018**). "Comparing Speech and Keyboard Text Entry for Short Messages in Two Languages on Touchscreen Phones," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies **1**(4), 1–23, doi: 10.1145/3161187.

Scott, A. J., Webb, T. L., Martyn-St James, M., Rowse, G., et Weich, S. (**2021**). "Improving sleep quality leads to better mental health : A meta-analysis of randomised controlled trials," Sleep Medicine Reviews **60**, 101556, doi: 10.1016/j.smrv.2021.101556.

Shen, J., Barbera, J., et Shapiro, C. M. (**2006**). "Distinguishing sleepiness and fatigue : focus on definition and measurement," Sleep Medicine Reviews **10**(1), 63–76, doi: 10.1016/j.smrv.2005.05.004.

Sorensen, K. (**2019**). "Millennials' Acceptance of Voice Activated Shopping," Textiles, Merchandising and Fashion Design : Dissertations, Theses, & Student Research .

The IQVIA Institute for Human Data Science (**2021**). "Digital Health Trends 2021," Institute Report.

Topol, E. (**2019**). *Deep Medicine : How Artificial Intelligence Can Make Healthcare Human Again* (Hachette UK).

Young, T. B. (**2004**). "Epidemiology of daytime sleepiness : definitions, symptomatology, and prevalence," The Journal of Clinical Psychiatry **65 Suppl 16**, 12–16.

Première partie

La voix comme mesure de pathologies

Résumé

Dans cette partie, nous proposons une première introduction de l'utilisation de la voix comme marqueur de pathologies. À partir d'une brève description du mécanisme de production de la voix dans le chapitre 1, nous formulons l'hypothèse que la production vocale est un marqueur sensible à de nombreuses pathologies de diverses natures.

Dans le chapitre 2, nous proposons une revue des différentes pathologies détectables dans la voix. Plus précisément, nous proposons une revue des corpus utilisés pour chacune des pathologies identifiées, dont nous discutons ensuite les méthodologies.

Mots-clés

Voix ; Modèle de production de la parole ; Revue de la littérature ; Diagnostic ; Méthodologie de construction de corpus

Chapitre 1

Qu'est-ce que la parole ?

Sommaire

1.1 Motivations

La conception de marqueurs vocaux permettant l'estimation de pathologies ou de symptômes nécessite la connaissance et la compréhension des mécanismes permettant la production de la voix et de la parole. En effet, en identifiant les mécanismes avec lesquels ces pathologies interfèrent, il est alors possible de concevoir des marqueurs spécifiques de cette altération de la production vocale.

Dans ce chapitre introductif, nous proposons une description des processus musculaires (section 1.2) et neurolinguistiques (section 1.3) impliqués dans la production de parole. Les mécanismes décrits seront notamment nécessaires à l'élaboration des marqueurs décrits dans la partie IV.

Enfin, dans la section 1.4, nous formulons l'hypothèse que la voix et la parole sont des mesures pertinentes de certaines pathologies et étudions les implications d'une telle hypothèse.

1.2 Production physique de la voix

1.2.1 Lien entre voix et respiration

La production d'une voyelle soutenue (tâche de vocalisation la plus simple) mobilise plus d'une centaine de muscles (Denes et Pinson, 1963). Les premiers sont les muscles respiratoires (diaphragme et muscles intercostaux) qui compressent les poumons afin de créer un flux d'air dans la trachée. Ce flux d'air passe ensuite à travers les plis vocaux qui, suivant s'ils sont tendus ou non, rentreront en résonnance. Le flux d'air arrive ensuite dans la cavité bucco-nasale où muscles de la langue, des joues et de la gorge amplifient et modulent le son. Ces ensembles musculaires sont représentés dans la figure 1.1.

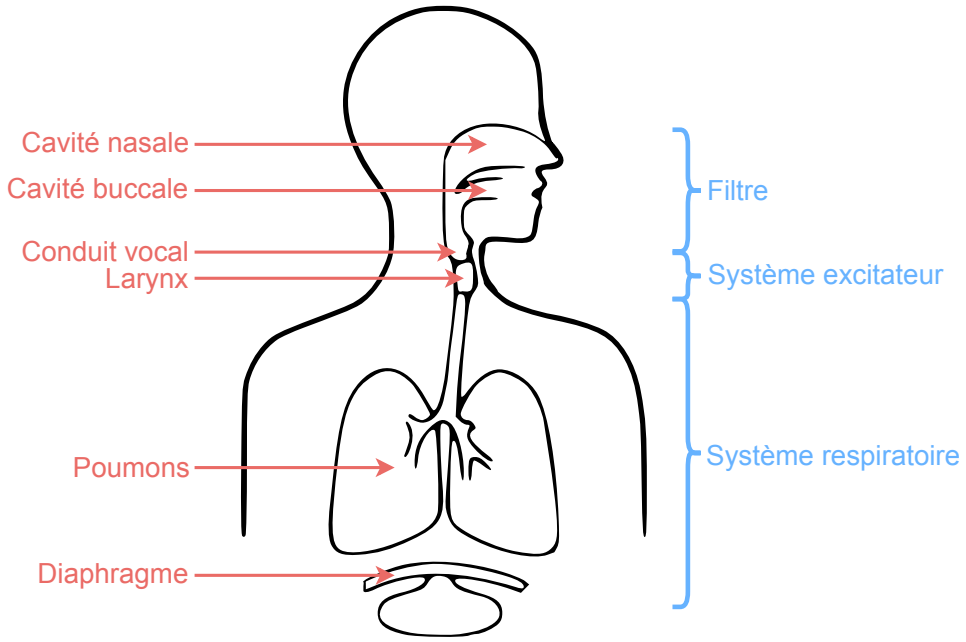

FIGURE 1.1 – Schéma des différentes parties de l'appareil respiratoire et phonatoire impliquées dans la production vocale.

1.2.2 Modèle source-filtre d'un son voisé

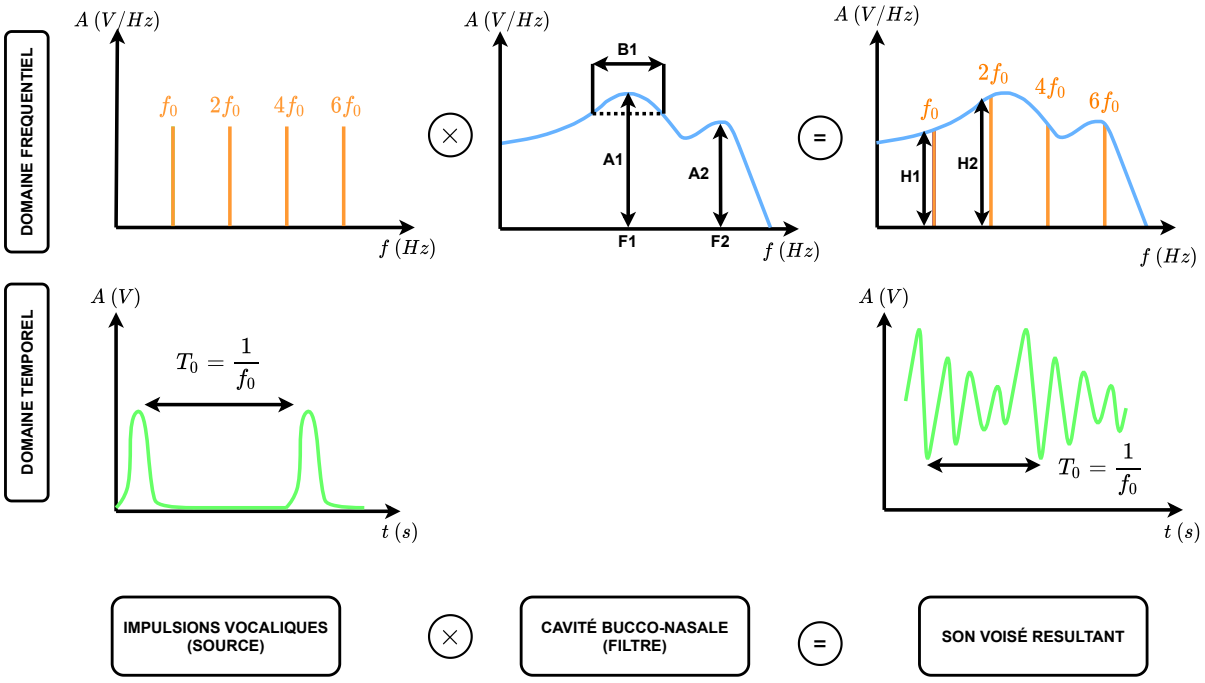

FIGURE 1.2 – Harmoniques et formants du signal vocal à partir d'un modèle source-filtre. Figure inspirée de (Denes et Pinson, 1963).

Cette activité motrice est habituellement expliquée sous le prisme d'un modèle source-filtre (Fant, 1970; Chiba et Kajiyama, 1958), pour lequel le signal vocal résulte de la résonance des impulsions initiées par les plis vocaux de la trachée dans la cavité bucco-nasale (cf. figure 1.2) :

— Les plis vocaux génèrent des impulsions glottiques espacées de T_0, correspondant dans le domaine fréquentiel à une suite d'impulsions de mêmes amplitudes et de fréquences multiples de $f_0 = \frac{1}{T_0}$, appelée *harmoniques* ;
— La cavité bucco-nasale correspond à un filtre, ayant une ou plusieurs fréquences de résonnance (notées F_i), d'amplitudes respectives A_i et de bandes passantes B_i. Ces fréquences de résonnances sont appelées *formants* ;
— Les impulsions glottiques sont filtrées par la cavité bucco-nasale, produisant ainsi un son *voisé*, ayant des caractéristiques fréquentielles liées à la fois aux harmoniques et aux formants.

Ce sont des statistiques calculées sur ces caractéristiques – les *harmoniques* et les *formants* – qui sont introduites dans le chapitre 11 comme premiers descripteurs vocaux de la somnolence.

Cette activité motrice s'intègre cependant dans un schéma plus large de production de la parole, notamment en interaction avec un ensemble de fonctions cognitives et neurolinguistiques qui perçoivent, analysent, et génèrent les différents signaux nécessaires à la production d'un signal de parole.

1.3 Fonctions neurolinguistiques

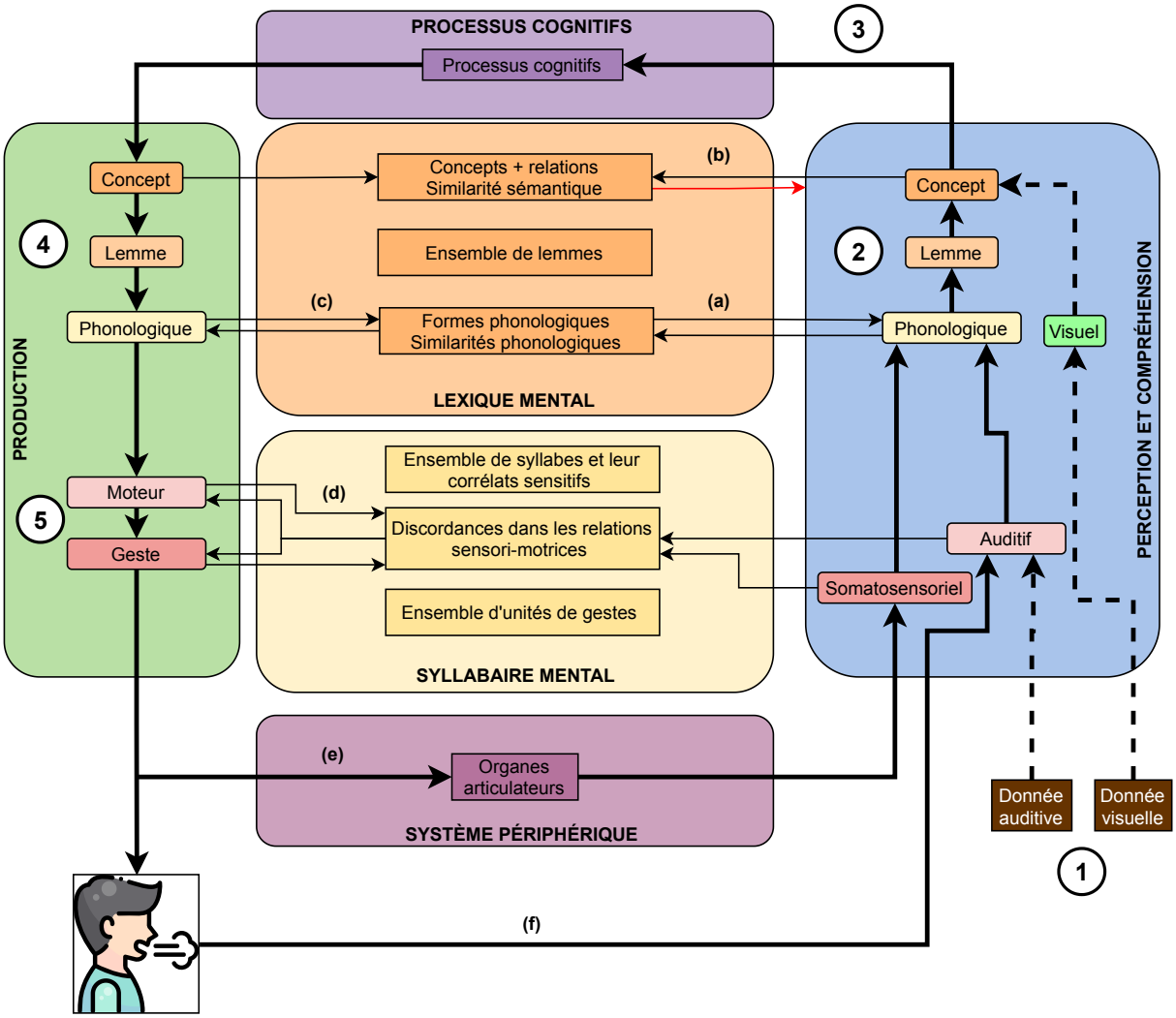

FIGURE 1.3 – Modèle de production de la parole proposé dans (Kröger *et coll.*, 2020).

Afin d'expliciter les fonctions neurolinguistiques impliquées dans la production de la parole, nous nous appuyons ici sur le modèle introduit par Kröger *et coll.* (2020), qui a été élaboré à partir d'un modèle génératif implémenté avec la librairie Python `Nengo` et validé sur différentes pathologies orthophoniques. Celui-ci est représenté dans la figure 1.3.

À partir de l'exposition du sujet à un stimulus externe (noté 1 sur le schéma), les processus neuronaux impliqués dans la réponse vocale sont les suivants :

1. Des stimulations auditives et/ou visuelles sont présentées au sujet.
2. Celles-ci sont d'abord traitées par un bloc de fonctions de perception et de compréhension du signal entrant (en bleu sur le schéma). Pour un signal visuel (par exemple la lecture d'un texte ou l'analyse d'une image), une fonction dédiée permettant de traiter l'information et d'en extraire un concept n'est pas détaillée ici. Pour un signal audio, celui-ci est décomposé en trois unités correspondant à celles utilisées dans le lexique mental : les phonèmes, les lemmes et enfin le concept. Le lexique mental contient

l'ensemble des formes phonologiques, lemmes et concepts et relations entre concepts connus du sujet. Ces connaissances permettent deux boucles de contrôle : un contrôle de la forme phonologique, en prenant en compte les similarités phonologiques (a), et un contrôle au niveau sémantique, qui en fonction du concept compris ou produit, influence l'intégralité du bloc de perception et de compréhension (b) ;

3. Le ou les concepts ainsi extraits sont ensuite traités par des processus cognitifs, qui déterminent la réponse à produire au(x) concept(s) compris ;
4. Celle-ci est codée, à partir du concept, en lemmes puis en phonèmes, avec le même contrôle phonologique que lors du décodage (c) ;
5. Cette chaîne phonologique est ensuite convertie en commandes motrices, assemblées en un geste, conduisant à la production vocale. Ces deux unités motrices sont rétrocontrôlées par le syllabaire mental. Celui contient l'ensemble des unités de gestes ainsi que l'ensemble des syllabes et de leurs corrélats sensitifs connus du sujet. Le syllabaire permet ainsi, en contrôlant les commandes motrices et de gestes, de réguler la production vocale (d) grâce à deux boucles de rétroactions :
   — une boucle auditive, qui permet de produire la cible phonologique grâce au système de perception auditive (f) ;
   — une boucle somatosensorielle à travers les organes articulateurs, qui permet d'évaluer la concordance entre la cible et la production à un niveau gestuel (e).

Ainsi, la production de la parole fait intervenir de très nombreux mécanismes, intervenants à différents niveaux conceptuels. Et si l'interférence produite par certaines pathologies avec ces mécanismes était caractéristique de celles-ci ?

1.4 La parole comme outils de mesure de pathologies

1.4.1 Différentes interférences

Au regard des très nombreuses fonctions qu'elle fait intervenir, la parole semble ainsi un outil très sensible aux phénomènes qui pourraient interférer avec elles. C'est le cas des pathologies de l'appareil phonatoire et de la cavité buccale (cancers orolaryngés, pathologies des cordes vocales ...) mais aussi des pathologies neurodégénératives, qui peuvent perturber le bon fonctionnement des boucles de production et de perception précédemment décrites.

Une autre catégorie de pathologies ou d'altérations d'états auxquelles la production de parole est sensible comprend les troubles intervenant sur les processus cognitifs, qui font le lien entre la compréhension et la génération de la réponse vocale adéquate aux stimulus. Ceux-ci comprennent les émotions, la fatigue, le stress, mais aussi différents troubles mentaux (Low *et coll.*, 2020).

1.4.2 Mesure de phénomène cognitif

Dans ce dernier cas, nous proposons de clarifier la notion de *mesure de phénomène cognitif*, dans le cadre de l'utilisation de marqueurs vocaux. En effet, les phénomènes cognitifs ont lieu dans le domaine de la cognition, tandis que les mesures se font sur des manifestations physiologiques de ces états cognitifs (Claverie, 2021). L'exemple de l'influence de l'état de somnolence sur les capacités de lecture sujet est illustré dans la figure 1.4.

Les erreurs de lecture (décrites précisément dans le chapitre 13) sont des mesures de la capacité de lecture du sujet, tandis que la latence d'endormissement au Test Itératif de Latence d'Endormissement est une mesure de l'état de somnolence du sujet. Ainsi, utiliser les

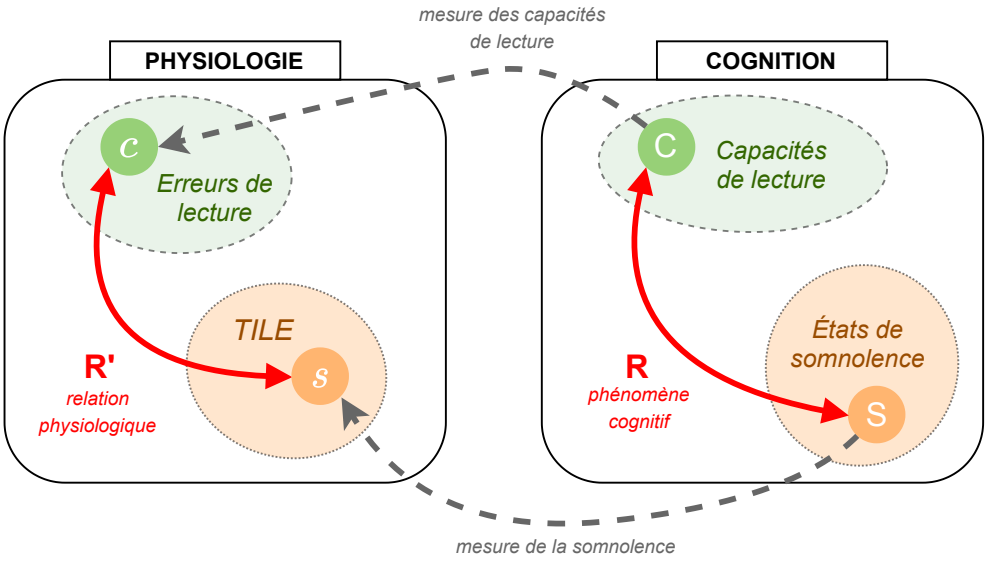

FIGURE 1.4 – Modèle de la mesure du niveau de somnolence, opérationnalisée par la latence d'endormissement lors d'un TILE, grâce aux capacités de lecture du locuteur, opérationnalisées par les erreurs de lecture. C : état des capacités de lecture du locuteur, S : état de somnolence du locuteur, c : état des erreurs de lecture du locuteur, s : latence d'endormissement au Test itératif de latence d'endormissement. Adapté de (Claverie, 2021).

erreurs de lecture pour détecter l'état de somnolence revient à supposer que le phénomène cognitif R « influence de l'état de somnolence sur les capacités de lecture » et son équivalent physiologique R' « influence de la latence d'endormissement sur la production des erreurs de lecture » sont similaires.

1.5 Conclusion

Nous avons présenté dans ce chapitre le fonctionnement de la production de la parole, ainsi que les différents processus qui peuvent être altérés par des pathologies ou des altérations d'état. Les changements observés dans la voix et dans la parole sont ainsi supposés être des mesures de l'altération des mécanismes spécifiquement altérés par ces pathologies, et permettent ainsi de les détecter.

Nous proposons dans le chapitre suivant une revue des différentes pathologies qui se sont avérées détectables grâce à des marqueurs extraits de la voix, ainsi que la méthodologie des bases de données utilisées par ces systèmes.

Chapitre 2

La parole comme outil diagnostique ou pronostic de pathologies

Sommaire

2.1 Contexte et méthode

Comme mentionné dans le chapitre précédent, les changements observés dans la voix et la parole peuvent être des marqueurs des altérations des mécanismes de leur production, elles-mêmes dues à des pathologies. Dans ce chapitre, nous proposons une revue des différentes pathologies qu'il est possible de diagnostiquer à partir de marqueurs vocaux.

Afin de constituer une liste des pathologies pour lesquelles des articles sont publiés (et donc sur lesquelles des équipes de recherche travaillent), nous avons appliqué une méthode *agrégative* : nous sommes partis tout d'abord de deux revues particulièrement complètes (Low *et coll.*, 2020; Fagherazzi *et coll.*, 2021), que nous avons enrichies au fur et à mesure de pathologies mentionnées dans les articles lus.

Nous avons ensuite catégorisé ces pathologies suivant l'élément de la chaîne de parole avec lesquelles elles interfèrent. Une attention particulière est portée aux corpus utilisés dans les études rapportées : les performances des systèmes présentés dans ces études sont toujours à mettre au regard d'une population ayant une certaine taille (qui doit être suffisante pour permettre une certaine généralisation), et certaines caractéristiques – notamment en termes de diagnostic et d'expression des symptômes.

Afin de limiter notre recherche et de décrire des résultats qui représentent les intérêts actuels de la communauté du traitement du signal de parole pour la détection de pathologies, nous nous limitons aux articles publiés entre le 1er janvier 2019 et le 31 décembre 2021. Sauf mention contraire, les performances de classification rapportées sont les taux de classification correcte (*accuracy*, cf chapitre 16) sur une tâche de classification binaire. Certaines études ne proposaient pas de résultats de classification ou de régression, mais une étude statistique sur le corpus décrit : dans ce dernier cas, aucune performance n'est rapportée.

De manière cohérente avec le chapitre précédent, nous commençons par décrire les systèmes identifiant les pathologies des cordes vocales et de la cavité bucco-nasale dans la section 2.2 et de certaines pathologies affectant le système respiratoire et cardiaque dans la section 2.3. Nous présentons dans la section 2.4 les systèmes détectant les pathologies affectant les capacités musculaires des sujets, et dans la section 2.5 les pathologies affectant les systèmes neurologiques. Dans la section 2.6, nous rapportons les corpus utilisés pour la détection de troubles psychiatriques, et dans la section 2.7 nous présentons les états altérant le fonctionnement général des locuteurs et locutrices. Enfin, dans la section 2.8, nous discutons les tendances générales précédemment observées sur les corpus utilisés par les études incluses dans la revue.

2.2 Pathologies des cordes vocales et de cavité bucco-nasale

Nous commençons cette revue par les pathologies touchant directement l'organe phonateur : les pathologies touchant les cordes vocales et la cavité bucco-nasale.

2.2.1 Hypothèse

Ces pathologies affectent directement les caractéristiques des plis vocaux (excitateur dans le modèle décrit dans le chapitre précédent) ou des différentes structures (muscles, cartilages, muqueuses) influençant la cavité bucco-nasale , qui joue le rôle de résonateur dans le modèle source-filtre décrit dans le chapitre précédent. Ces dégradations de la voix dues aux altérations de l'excitateur et/ou du résonateur sont caractéristiques de ces pathologies.

2.2.2 Troubles de la voix, dysarthrie et cancers ORL

Saarbruecken Voice Database La base de donnée la plus utilisée pour la détection de pathologies affectant la voix est la Saarbruecken Voice Database [SVD, (Woldert-Jokisz, 2007)]. Cette base de donnée, enregistrée en Allemagne, contient les enregistrements de 687 sujets sains et de 1354 patients atteints de plus de 70 pathologies différentes (par exemple différents types de dysphonie, laryngite, rhinophonie, aphonie ...).
Deux tâches sont réalisées par les sujets :

— d'une part, la vocalisation de voyelles soutenues (/a/, /u/, /i/) ;
— d'autre part la lecture à voix haute d'une phrase en allemand : « Guten Morgen,wie geht es Ihnen ? »

Les performances obtenues sur différents sous-corpus de cette base de données contenant différentes combinaisons de pathologies sont aux alentours de 90% pour le diagnostic (AL-Dhief *et coll.*, 2021; Chui *et coll.*, 2020; Danilovaité, 2021; Gidaye *et coll.*, 2020; Hammami *et coll.*, 2020; Mohammed *et coll.*, 2020; Syed *et coll.*, 2021b,a) et le diagnostic différentiel de différentes pathologies (AL-Dhief *et coll.*, 2021; Hammami *et coll.*, 2020; Syed *et coll.*, 2021b).

VOICED Un deuxième corpus utilisé massivement est le corpus Voice ICar fEDerico II Database [VOICED, (Cesari *et coll.*, 2018)]. Ce corpus, enregistré en Italie, contient les enregistrements de 58 sujets sains et de 150 patients atteints de dysphonies et de reflux laryngo-pharyngé, diagnostiqués par un ORL. Dans ce corpus, une seule tâche est proposée aux sujets, qui doivent vocaliser la voyelle /a/ pendant 5 secondes.

Les performances obtenues sur ce corpus sont également très hautes : autour de 90% dans (Chui *et coll.*, 2020), 98.6% dans (Chen *et coll.*, 2021).

Massaschussetts Eye and Ear Infirmary – MEEI Dans le Massaschussetts Eye and Ear Infirmary corpus [MEEI, (Voice and Speech Lab)], 53 sujets sains et 77 sujets atteints de paralysie des cordes vocales, de polypes ou de kystes dans la gorge ont été enregistrés en train de vocaliser la voyelle /a/ pendant 5 secondes et en train de lire le texte *Rainbow Passage*. Ce corpus a été enregistré aux États-Unis.

Arabic Voice Pathology Database – AVDP Cette base de données, enregistrée en Arabie Saoudite, contient les enregistrements de 350 sujets (118 sujets sains et 232 sujets pathologiques), diagnostiqués par un ORL et en comparaison avant/après chirurgie (Mesallam *et coll.*, 2017). Ces sujets ont été enregistrés en train de vocaliser des voyelles soutenues (/a/, /u/, /i/), de lire des mots isolés, en train de compter de 0 à 10 ou encore sur une tâche de parole spontanée non précisée.

Principe de Asturias – PdA Ce corpus, enregistré en Espagne par Arias-Londoño *et coll.* (2011), regroupe les enregistrements de 440 sujets différents (239 sujets sains, et 201 patients de 15 étiologies organiques et traumatiques différentes), enregistrés en train de vocaliser la voyelle /a/ pendant 3 secondes.

L'étude proposé par Gidaye *et coll.* (2020) portant à la fois sur le MEEI, l'AVPD et le PdA, a atteint une aire sous la courbe ROC de 1.0 pour la tâche de classification diagnostic (sujets sains vs sujets pathologiques) avec des marqueurs basés sur des ondelettes, tandis qu'une autre étude basée sur des sous-corpus de l'AVDP, du MEEI, du PdA et du SVD (regroupant 3995 échantillons enregistrés par des sujets sains et 4047 échantillons enregistrés par des patients) a atteint un F1-score de 73.3% (Harar *et coll.*, 2020).

UEX Ce corpus, également enregistré en Espagne, contient les enregistrements de 30 sujets sains et de 84 patients atteints de trois pathologies différentes (nodules, polypes ou Oedème de Reinke), diagnostiqués par un examen ORL, et enregistrés en train de vocaliser la voyelle /a/ pendant 5 secondes (Carrón *et coll.*, 2021).

Une étude récente, basée sur les corpus SVD, MEEI, PdA et UEX a atteint des scores de classification supérieurs à 95% (Madruga *et coll.*, 2021).

Torgo Cette base de données, bien que ne contenant les enregistrements que de huit patients évalués avec la *Frenchay Dysarthria Assessment*, a l'originalité d'avoir enregistré les patients sur une très grande variété de tâches différentes : voyelles soutenues, tests diadococinésiques (incluant /pa/-/ta/-/ka/), répétition de nombre de 1 à 10, alphabet international, lecture de texte (*Grandfather passage*), lecture de plus de 400 mots isolés et de plus de 500 phrases issues de tests d'orthophonie, et la description de 30 images tirées des *Webber Photo Cards*.

Associé avec l'UEX et le PdA, une étude a atteint sur ce corpus une performance de classification 85.12% (Narendra et Alku, 2020).

C2SI Ce corpus, d'initiative française, contient les enregistrements de 40 sujets sains et 87 patients remis d'un cancer de la cavité buccale et/ou du pharynx, enregistrés 6 mois après la fin du traitement (Woisard *et coll.*, 2021). Une des originalités de ce corpus repose sur les tâches qui sont proposées, qui ont été conçues pour évaluer l'ensemble du spectre de la communication parlée : voyelles soutenues (/a/), lecture de 52 pseudo-mots, 50 phrases nécessitant le décodage de leur sens, une lecture de texte (*La chèvre de M. Seguin* d'A. Daudet), ainsi que trois tâches phonétiques (lecture de phrases avec intonations différentes, oppositions paradigmatiques et syntactiques.). Ces tâches sont complétées par la description d'une image et une tâche de parole spontanée (le sujet donne son avis sur le questionnaire qu'il a rempli en début de session).

Autres corpus Les trois autres corpus que nous avons identifiés sont un corpus brésilien (Alves *et coll.*, 2021), un corpus Indien (Gour *et coll.*, 2020) et un corpus chinois (Hu *et coll.*, 2021) contenant respectivement 20, 55 et 189 sujets sains et 45, 55 et 552 sujets pathologiques. Les sujets sont enregistrés sur des voyelles soutenues (/a/) dans les trois corpus, et le troisième (Hu *et coll.*, 2021) propose également une tâche de lecture d'un texte, en mandarin. Les performances de classification sur ces trois corpus sont excellentes : 100% de classification correcte pour chaque pathologie vs sujets sains dans (Alves *et coll.*, 2021), 76.56% de classification correcte des sujets sains vs sujets pathologiques dans (Gour *et coll.*, 2020) et 98% d'AUC sur deux classes et 85% d'AUC sur cinq classes dans (Hu *et coll.*, 2021).

2.3 Pathologies respiratoires et cardiaques

2.3.1 Hypothèse

La parole est principalement constituée de sons produits lors de l'expiration d'air provenant des poumons. Par leurs interférences caractéristiques avec les fonctions respiratoires, certaines pathologies pulmonaires peuvent donc être identifiées à travers des signaux de parole. De plus, les fonctionnements pulmonaires et cardiaques sont étroitement liés : par leur impact sur le fonctionnement respiratoire, certaines pathologies cardiaques semblent être détectables grâce à des marqueurs vocaux.

2.3.2 Pathologies cardiaques

Diabète

Une des premières pathologies cardiaques que nous avons identifiées, mentionnée dans (Fagherazzi *et coll.*, 2021), est le diabète.

Dans les travaux de Pinyopodjanard *et coll.* (2021), en Thaïlande, une étude statistique a trouvé des caractéristiques différenciant 83 sujets diabétiques de 70 sujets sains, enregistrés en train de vocaliser une voyelle /a/ soutenue.

Pathologies cardiaques chroniques

Nous également trouvé mention de détection de pathologies cardiaques chroniques, notamment avec l'étude menée par Bourouhou *et coll.* (2021) au Maroc. Les 40 sujets sains et 35 patients d'un service de cardiologie atteints de différentes pathologies cardiaques chroniques ont été différenciés avec un score de classification correcte de 81.51%, sur des enregistrements des voyelles /a/, /o/, et /i/ effectués avec un smartphone dans les locaux de l'hôpital.

Crise cardiaque

Enfin, non plus dans une perspective de détection de pathologie chronique, mais dans une volonté de détection de crise, l'équipe finnoise de (Kiran Reddy *et coll.*, 2021) a différencié, avec une performance 81.51%, les enregistrements faits de 25 sujets sains et de 25 patients adressés aux urgences pour crise cardiaque, indépendamment de l'étiologie sous-jacente. Les enregistrements ont été effectués à l'hôpital, sur la lecture d'un texte de 80 mots et lors d'un entretien clinique.

2.3.3 Covid-19

Lors de l'émergence de la pandémie de Covid-19 en 2020, un effort de recherche inédit a été mis en place pour constituer des bases de données et élaborer des solutions de détection de la Covid-19 dans la voix. Deux grands types de corpus ont vu le jour : des corpus 'traditionnels', contenant les enregistrements d'au plus quelques dizaines de patients et de sujets sains, enregistrés dans des conditions contrôlées et dont le statut pathologique est certifié par un examen ou un test médical ; et d'autre part des corpus collectés sur des volontaires, partout autour de la planète, grâce à des applications smartphones. Si ces derniers permettent de collecter de grandes quantités de données à moindre coût, l'hétérogénéité des langues, profils, cultures et la non-vérification du statut pathologique freine grandement la reconnaissance des systèmes élaborés par ces systèmes par les cliniciens.

Dans cette revue, nous ne considérons pas les corpus contenant uniquement des enregistrements de toux et nous rapportons, sauf mention contraire, les performances obtenues uniquement sur le sous-corpus d'échantillons vocaux.

Approches écologiques

Coswara Un des deux plus grands corpus existant sur la collecte de voix en situations réelles pour la détection de la Covid-19 dans la voix est l'initiative indienne Coswara (Sharma *et coll.*, 2020). Cette application mobile, lancée dès le début de la crise, demande aux volontaires de soutenir les voyelles /a/, /e/ et /o/, ainsi que de compter à voix haute, à vitesse normale

et rapidement. La vérité terrain de cette base de données est très peu robuste : les sujets indiquent dans l'application s'ils sont sains, exposés, soignés ou infectés par la Covid-19.

Sur ce corpus, en constante évolution, les travaux de Verde *et coll.* (2021b,a) ont atteint des scores de classification de 82.35% (sur un sous-corpus de 83 Covid-/83 Covid+) et 97.1% (sur un sous corpus contenant 950 Covid- et 81 Covid+). Dans (Dash *et coll.*, 2021), un score de 73.5% a été atteint sur la tâche de comptage rapide, sur un sous-corpus de Coswara contenant 570 participants.

Cambridge dataset Le Cambridge dataset est l'autre grand corpus participatif sur la Covid-19 (Xia *et coll.*, 2021). Dans ce corpus, collecté en conditions écologiques grâce à une application smartphone, les sujets doivent dire la phrase « J'espère que mes données vont aider à la gestion de la pandémie », dans la langue de leur choix. La vérité terrain de ce corpus est plus robuste que dans Coswara : les utilisateurs peuvent indiquer s'ils ont été diagnostiqués par un test PCR, un médecin, ou s'ils se sont autodiagnostiqués. De plus, ceux-ci peuvent indiquer s'ils ont des symptômes liés à l'infection, et les lister.

De même que pour Coswara, un des défis dans le rapport du contenu de ce type de bases de données est l'hétérogénéité des profils et langues des utilisateurs, en plus d'avoir un corpus qui est en constant changement.

Sur ce corpus, Han *et coll.* (2021) proposent différents diagnostics (infecté/sain, avec/sans symptômes) ayant des AUC de l'ordre de 0.65-0.77. L'article publié par Mendonça *et coll.* (2021) rapporte un score de classification de 75.5% sur un sous-corpus du Cambridge dataset tandis que Xia *et coll.* (2021) fait état d'une AUC de 0.61 sur un autre sous-corpus.

audEERING GmbH et University of Augsburg L'étude proposée dans (Hecker *et coll.*, 2021) repose sur deux corpus à la méthodologie très semblable, basée sur l'enregistrement de sujets allemands sur les deux premières phrases de *La bise et le soleil* et sur les voyelles soutenues /a/, /e/, /i/, et /u/.

La différence entre les deux corpus est le lieu des enregistrements : alors que le premier est enregistré en conditions écologiques, avec pour vérité terrain la seule déclaration des participants, le deuxième est collecté dans les locaux de l'université d'Augsburg, et les sujets sont enregistrés dans les trois jours suivant un test.

Sur les 13 patients infectés (7 dans audEERING GmbH et 6 dans le corpus de l'université d'Augsburg) et les 26 sujets sains (respectivement 14 et 12) inclus, le système élaboré par Hecker *et coll.* (2021) atteint des scores pour différentes combinaisons de diagnostics de l'ordre de 63%.

Enregistrements en conditions contrôlées

SondeHealth Enregistré dans un hôpital aux États-Unis, le SondeHealth corpus (Sonde One, 2021) contient les enregistrements de 22 patients infectés par le virus (déclaration de test), 22 sujets sains avec symptômes et de 22 sujets sains sans symptômes. Ceux-ci ont été enregistrés grâce à des smartphones, en train de vocaliser la voyelle /a/, de prononcer /pa/-/ta/-/ka/ et en train de lire à voix haute la phrase "Mama made some lemmon jam".

Les performances proposées dans (Stasak *et coll.*, 2021d) sur différentes tâches de classification (infecté/sain, symptômes/sans symptômes) avoisinent les 82-84%.

Chili Covid Dataset Enregistré au Chili, ce corpus contient les enregistrements de 9 patients infectés et de 10 sujets sains, tous enregistrés dans les sept jours suivant un test, sur des

voyelles soutenues (/a/, /i/, /u/). Les travaux reportés par Deshmukh *et coll.* (2021) font état d'une AUC de 0.90 tandis que ceux présentés dans (Al Ismail *et coll.*, 2021) rapportent une AUC de 0.912 sur la seule voyelle /i/.

Pinkas et col. 2020 Dans (Pinkas *et coll.*, 2020), le corpus contient les enregistrements de 29 patients infectés et de 59 sujets sains ayant été testés (test PCR) lors de la session d'enregistrement, durant laquelle ils ont vocalisé les phonèmes /a/ et /z/, et compté de 50 à 80. L'article introduisant le corpus fait état de performances de classification (F1-score) de l'ordre de 0.81 sur le phonème /z/, 0.74 sur le phonème /a/ et de 0.80 sur la tâche de comptage.

Bartl-Pokorny et col. 2021 De même, les sujets allemands inclus dans le corpus introduit par Bartl-Pokorny *et coll.* (2021) contient les enregistrements de 11 patients infectés et de 11 sujets sains, tous testés et enregistrés dans les 3 jours après un test. Ceux-ci ont été enregistrés avec un smartphone, sur les voyelles /a/, /e/, /i/, /o/ et /u/.

Shimon et col. 2021 Dans (Shimon *et coll.*, 2021), le corpus, enregistré en Israël, contient les enregistrements de 130 patients infectés et de 69 sujets sains, tous « vérifiés médicalement », en train de vocaliser les phonèmes /a/, /s/ et /z/. Le système proposé permet d'atteindre un score de classification de 78%.

Asiaee et col. 2020 et Sondhi et col. 2021 Dans (Asiaee *et coll.*, 2020) et (Sondhi *et coll.*, 2021), les paradigmes d'évaluation des sujets pathologiques sont robustes (test PCR + CT scan pour le premier, test PCR pour le deuxième), mais pas la définition des sujets sains (pas de contact avec covid-19+ et pas de voyage pour le premier, pas de problème de santé et jamais diagnostiqué covid-19+ pour le deuxième). Ainsi, le premier contient les enregistrements de 64 patients infectés et de 70 sujets sains, enregistrés en Iran sur la voyelle soutenue /a/. Dans le deuxième corpus, ce sont 16 patients Indonésiens avant et après infection et 20 sujets sains qui ont été enregistrés avec un smartphone sur les voyelles /a/, /e/, /i/, /o/ et /u/.

Suppakitjanusant et col. 2021 La tâche de classification proposée dans (Suppakitjanusant *et coll.*, 2021) est légèrement différente des tâches précédentes, puisqu'il s'agit d'identifier non pas les patients infectés par le Covid, mais ceux l'ayant été dans les précédentes semaines. Ce corpus contient les enregistrements de 76 sujets post-infection (8 semaines après l'infection) et de 40 sujets sains, sur la voyelle /a/ soutenue et un texte polysyllabique sélectionné par un ORL. Sur la lecture de texte, l'algorithme proposé permet d'atteindre un score de classification de 85%.

Quatieri et col. 2020 Enfin, le travail proposé par Quatieri *et coll.* (2020) diffère de toutes les méthodologies précédentes. En effet, le corpus proposé repose sur les conférences de presse et interviews diffusées à la télévision de personnes avant et après avoir été infectées, sans symptômes évidents, collectés sur YouTube, Twitter et Instagram.

2.3.4 Autres pathologies pulmonaires

En dehors de l'énorme effort de la communauté dans l'identification de l'infection à travers la voix, nous avons également trouvé des travaux sur d'autres pathologies affectant les voies respiratoires.

Asthme

B.T. et col. 2020 Une première pathologie que nous avons identifiée dans la littérature est l'asthme. Dans (B. T. *et coll.*, 2020), 71 patients Singapouriens asthmatiques ont été recrutés à l'hôpital en parallèle de 135 sujets sains et enregistrés grâce à des smartphones en train de soutenir la voyelle /a/. L'algorithme conçu par l'équipe a obtenu une performance de classification de 72.2%.

Yadav et col. 2020 Dans (Yadav *et coll.*, 2020), 48 patients Bengalais et 48 sujets sains ont été enregistrés en train de soutenir les phonèmes /a/, /i/, /u/, /eɪ/, /oʊ/, /s/, /z/. Une originalité du protocole proposé dans cette étude est l'utilisation d'une pince à linge sur le nez afin de concentrer le flux d'air dans la cavité buccale. L'étude atteint, sur le phonème /oʊ/, 75.4% de classification correcte.

Hypertension pulmonaire – HP

Aux Pays-Bas, van Bemmel *et coll.* (2021) proposent la différenciation entre patients lors d'une crise, patients stabilisés et sujets sains, enregistrés en train de lire un texte (*de König*) coupé en deux : les patients en crise lisent le début du texte, puis lisent la fin du texte une fois stabilisés. L'algorithme proposé permet d'atteindre un score de 100% de Corrélation de Matthews.

Bronchopneumopathie chronique obstructive – BCO

Sara *et coll.* (2020) proposent un corpus constitué de 27 sujets sains et 56 patients atteints de BCO résultant de diverses pathologies. Ces sujets sont enregistrés avec un smartphone lors de la lecture d'un texte et la description d'une expérience positive/négative.

Maladie pulmonaire chronique – MPC

Enfin, Saleheen *et coll.* (2020) proposent un corpus contenant les enregistrements de 153 patients atteints de diverses maladies pulmonaires chroniques et de 58 sujets sains. Ceux-ci sont enregistrés sur une voyelle /a/ soutenue avec des smartphones. L'algorithme proposé par l'équipe permet d'atteindre un score de classification (F1-score) de 71% sur la détection d'une pathologie.

2.4 Altération des capacités musculaires – sclérose en plaques

Une troisième catégorie de troubles pouvant être détectés dans le signal de parole regroupe les pathologies interférant avec les capacités musculaires de manière générale, pouvant toucher à la fois les muscles respiratoires et ceux de la gorge et de la cavité bucco-nasale.

Nous avons identifié trois corpus permettant la détection de la sclérose en plaques à partir d'échantillons vocaux.

Fazeli et col. 2020 Dans (Fazeli *et coll.*, 2020), un corpus iranien contenant les enregistrements de 47 patients dont le diagnostic est confirmé par un neurologue et de 20 sujets sains est proposé. La tâche proposée est la vocalisation de la voyelle /a/ avec différentes variations : aussi longtemps que possible, confortablement, à la plus haute fréquence possible, aussi faiblement que possible.

Noffs et col. 2020 Dans le corpus australien proposé par Noffs *et coll.* (2020), 110 patients dont le diagnostic est confirmé par un neurologue et 22 sujets sains sont enregistrés sur différentes tâches : la voyelle /a/ soutenue (10s), les jours de la semaine dans l'ordre, /pa/-/ta/-/ka/, la lecture d'un paragraphe phonétiquement équilibré et la narration d'une histoire personnelle (1 min.).

Schultz et col. 2021 Enfin, dans le corpus australien proposé par Schultz *et coll.* (2021), 32 patients atteints de sclérose en plaques, 32 sujets sains et 32 sujets atteints d'ataxie de Friedreich, diagnostiqués par un neurologue, ont été enregistrés en train de lire un texte à voix haute (*Grandfather passage*).

2.5 Pathologies neurologiques

2.5.1 AVC/AIT

Les troubles neurologiques peuvent apparaître sous forme de crise, notamment lors d'Accidents Vasculaire Cérébraux ou d'Accidents Ischémiques Transitoires. Ces crises laissent de nombreuses séquelles, notamment sur la voix et la parole. Une revue de ces conséquences est proposée dans (Chiaramonte et Vecchio, 2021).

Nous avons trouvé dans la littérature deux corpus récents qui contiennent des enregistrements de patients ayant été sujets d'AVC ou d'AIT.

De Cock et col. 2021 Tout d'abord, dans (De Cock *et coll.*, 2021), un corpus regroupe les enregistrements 67 patients et de 84 sujets sains. Ils sont enregistrés lors de la passation du *Radbound Dysarthria Assessment* et de la *Dutch Sentence Intelligibility Assessment*. Les patients ont reçu un premier diagnostic d'AIT, et le type de la dysarthrie qui en résulte est classifié grâce au système de classification de Mayo (*Mayo Classification System*).

Min et col. 2020 Dans (Min *et coll.*, 2020), les échantillons de 46 patients et 173 sujets sains qui ont été enregistrés lors de la prononciation du mot coréen "TTOGTTAG" ont été différenciés par des analyses statistiques du *Voice Onset Time*.

2.5.2 Maladie de Parkinson

Par leur impact à la fois sur les capacités motrices, mais aussi cognitives, les maladies neurodégénératives peuvent également être détectées dans la voix et la parole. C'est le cas de la maladie de Parkinson, qui elle aussi peut être la cause de certaines formes de dysarthries et de dysphonies, et qui concentre une partie conséquente des efforts de la communauté du traitement du signal de parole.

Enregistrements en conditions contrôlées

PC-GITA (Orozco-Arroyave *et coll.*, 2014) Un des corpus les plus utilisés pour cette tâche est le PC-GITA, enregistré en Colombie. Celui-ci contient les enregistrements de 50 patients et de 50 sujets sains en train de réaliser les tâches suivantes : répétition de 25 mots isolés, lecture de 10 phrases, un texte, 5 voyelles soutenues et une tâche de parole spontanée. Les patients sont enregistrés durant la phase ON (<3h après médication). Il a notamment servi de corpus pour la compétition Interspeech 2015 sur la détection de Parkinson (Schuller *et coll.*, 2015).

Sur ce corpus, Kadiri et Alku (2020) atteignent un score de classification de 76.0%, tandis que les travaux proposés par Amato *et coll.* (2021) permettent d'atteindre 99.4% sur une version augmentée du PC-GITA (68 patients et 69 sujets sains).

UCI Un deuxième corpus, très utilisé historiquement dans la tâche, est le corpus UCI, aussi appelé parfois le *Max Little corpus* (Little *et coll.*, 2008). Ce corpus américain contient les enregistrements de 23 patients et de 8 sujets sains, en train de soutenir la voyelle /a/.

Les travaux récents sur ce corpus présentent des performances de classification très hautes, comme par exemple 96.70% dans (Asmae *et coll.*, 2020), 96.92% dans (Despotovic *et coll.*, 2020) ou encore 98.3% dans (Sajal *et coll.*, 2020).

PSD dataset Un troisième corpus très utilisé dans l'état de l'art est le PSD dataset, collecté en Turquie (Sakar *et coll.*, 2013). Ce corpus contient les enregistrements de 20 patients et de 20 sujets sains sur des tâches de lecture, de voyelles soutenues et d'énonciation de lettres et de chiffres.

Sur ce corpus, Rizvi *et coll.* (2020) atteignent un score de 99.03% de classification correcte, tandis que Zhang *et coll.* (2021) atteignent 96.54%.

UEX Une quatrième base de données très employée pour la tâche est la base de données UEX (Carrón *et coll.*, 2021). Cette base de données, enregistrée en Espagne, contient les enregistrements de 30 sujets sains et de 30 patients atteints de la maladie de Parkinson, ces derniers ayant un diagnostic final de la maladie. Les patients sont enregistrés, avec un smartphone en conditions contrôlées, en train de vocaliser la voyelle /a/.

Sur ce corpus, le système proposé par Carrón *et coll.* (2021) atteint 92% de classification correcte.

Approche écologique

mPower database Le pendant du précédent corpus en conditions écologiques est la mPower database (Bot *et coll.*, 2016). De même que pour l'UEX, les enregistrements contenus dans le corpus mPower sont effectués avec un smartphone sur des voyelles /a/ soutenues. En revanche, ceux-ci sont effectués en conditions écologiques. Comme la majorité des corpus collectés en conditions écologiques, sa taille est variable, et les études travaillant sur ce corpus reportent systématiquement des tailles de corpus différentes.

Par exemple, dans (Carrón *et coll.*, 2021), 71% de performances sont atteintes sur 30 patients et 30 sujets sains extraits de la base de données, alors que dans (Karaman *et coll.*, 2021), ce sont 23288 échantillons enregistrés par des sujets sains et 10589 échantillons enregistrés par des patients qui permettent la classification correcte à 89.75% de la pathologie.

Approche multilingue

Une des rares études multilingues de cette revue est menée dans (Vásquez-Correa *et coll.*, 2019), qui utilise trois corpus, dans trois langues différentes, pour comparer les performances de systèmes avec différentes combinaisons d'entraînements.

Sur les trois corpus, les tâches sélectionnées sont : /pa/-/ta/-/ka/, la lecture de phrases, d'un texte de 80 mots et un monologue d'une minute. Pour l'espagnol, le corpus utilisé est le PC-GITA, tandis que pour l'Allemand le corpus utilisé contient les enregistrements de 88

sujets sains et de 88 patients ; pour le Tchèque, le corpus utilisé contient 50 sujets sains, et 50 patients.

Sur ces trois corpus, les performances de classification binaires sont de l'ordre de 70%

Corpus en français

Jeancolas et col. 2021 Un corpus français est proposé par Jeancolas *et coll.* (2021), contenant les enregistrements de lecture de texte, dialogues, phrases, de parole spontanée, et de répétitions rapides de syllabes, de 115 patients et 91 sujets sains, enregistrés durant leur phase ON (moins de 12h après la prise du traitement). Une des faiblesses de la conception de ce corpus repose sur la qualité des enregistrements effectués : alors qu'une portion du corpus est enregistrée avec un microphone de haute qualité, l'autre partie est enregistrée au téléphone, en basse qualité.

MonPaGE Une autre initiative francophone pour l'étude de la maladie de Parkinson à mentionner est MonPaGE (Laganaro *et coll.*, 2021; Lévêque *et coll.*, 2016), un protocole de dépistages des troubles parkinsoniens en libre accès, contenant à la fois un protocole d'enregistrement de la parole et un logiciel d'analyse des caractéristiques vocales.

Corpus pour diagnostic différentiel

Trois corpus ont été identifiés comme permettant un diagnostic différentiel avec la maladie de Parkinson.

Rusz et col. 2021 Rusz *et coll.* (2021) ont introduit un corpus multinationalités (Autriche, Allemagne, Italie, France, États-Unis, Canada) contenant les enregistrements de 117 sujets sains, 109 patients atteints de la maladie de Parkinson et de 150 patients atteints de troubles du comportement en sommeil paradoxal (*REM sleep behavior disorder* – RBD) en train de vocaliser la voyelle /a/. La différenciation de patients avec maladie de Parkinson de ceux affectés par un RBD trouve son intérêt dans le lien entre les deux pathologies, la présence de troubles du comportement en sommeil paradoxal pouvant être un marqueur précoce de l'arrivée de la maladie de Parkinson (Postuma *et coll.*, 2006, 2010; Boeve, 2013).

Arora et col. 2021 Dans (Arora *et coll.*, 2021), un corpus étatsunien d'enregistrements de voyelles soutenues /a/ a été collecté avec des smartphones sur 335 patients atteint de la maladie de Parkinson, 92 sujets sains et 112 patients atteints de parasomnies en sommeil paradoxal. Le diagnostic différentiel atteint une sensibilité et une spécificité de l'ordre de 60%.

Mallela et col. 2020 Enfin, Mallela *et coll.* (2020) proposent un corpus contenant les enregistrements de 60 sujets contrôles, 60 patients atteints de Parkinson et de 60 patients atteints de sclérose amyotrophique latérale (Amyotrophic Lateral Sclerosis – ALS). Ces sujets indonésiens sont enregistrés lors de la description d'images (entre 40 et 60 images), lors de la production de phonèmes soutenus (/a/, /i/, /o/, /u/, /eo/,/s/, /sh/, /f/), lors de la prononciation rapide de /pa/-/ta/-/ka/ et lors d'une tâche de parole spontanée, où il leur est demandé de décrire un festival et un lieu où ils sont allés récemment. Les performances de classification des systèmes proposés dépassent les 80%.

Autres corpus et initiatives

Nous avons identifié 66 autres travaux sur la maladie de Parkinson, que nous ne détaillons pas ici.

2.5.3 Maladie d'Huntington

Une deuxième pathologie neurodégénérative qui suscite l'intérêt de la communauté est la maladie d'Huntington. Par rapport aux autres pathologies mentionnées dans ce chapitre, le diagnostic de celle-ci se fait par identification d'une mutation sur un gène, responsable de la maladie.

La maladie d'Huntington se traduit par des symptômes moteurs (chorée d'Huntington), cognitifs (perte de mémoire, difficulté à la planification) et éventuellement psychiatriques. Nous avons trouvé trois corpus dans la littérature pour la détection de la maladie de Huntington dans la voix.

Tovar et col. 2020 Dans (Tovar *et coll.*, 2020), 20 patients atteints de la maladie d'Huntington (10 avec symptômes, 10 sans symptômes) ont été enregistrés lors d'une entrevue de 15 minutes et lors de la description de deux vidéos, l'une avec dialogue et l'autre sans.

Romana et col. 2020 Dans (Romana *et coll.*, 2020), 31 sujets sains et 31 patients (12 sans symptômes, 11 en stade précoce et 7 en stade avancé) ont été enregistrés en train de lire un texte (*Grandfather passage*) et de vocaliser la voyelle /a/. L'algorithme proposé a permis de différencier les patients avec symptômes de ceux sans symptômes avec une performance de 87.0%.

Dong et col. 2020 Enfin, une étude chinoise s'est intéressée à l'effet de la chorée d'Huntington sur une langue non syllabique (le mandarin) (Dong *et coll.*, 2020), en étudiant le comportement de lecture de 5 patients et de 5 sujets sains passant une longue batterie de tests incluant notamment /pa/-/ta/-/ka/, et une tâche d'expression orale.

2.5.4 Maladie d'Alzheimer

Enfin, une troisième maladie neurodégénérative pouvant être détectée dans le signal de voix ou de parole est la maladie d'Alzheimer. Une revue de la détection de cette pathologie est proposée par (Pulido *et coll.*, 2020).

Enregistrements en conditions contrôlées

DementiaBank Dataset La plupart des corpus existant pour la détection d'Alzheimer sont dérivés du corpus étatsunien DementiaBank. La grande force de ce corpus est le diagnostic d'Alzheimer, qui repose sur au moins trois visites avec des neurologues, et même, pour certains patients, une autopsie confirmant le diagnostic.

Les corpus extraits du corpus DementiaBank sont marqués avec un astérisque dans les paragraphes suivants.

Pitt corpus* Ce corpus contient les enregistrements de 312 sujets (208 patients atteints de la maladie d'Alzheimer et 104 sujets sains), lors de la description d'une image (*Cookie Theft*), de tâches de fluence verbale, d'une tâche de construction de phrases et d'une tâche de mémoire (raconter une histoire) (Becker *et coll.*, 1994).

Les systèmes proposés dans (Haider *et coll.*, 2020; Liu *et coll.*, 2021; Pan *et coll.*, 2020; Roshanzamir *et coll.*, 2021) atteignent tous des performances de classification supérieures à 78% sur ce corpus.

Adress* Ce corpus, proposé pour la compétition Interspeech 2020 sur la détection de la maladie d'Alzeihmer (Luz *et coll.*, 2020), est un sous-corpus du corpus Pitt, contenant uniquement la tâche de description d'image de 78 sujets sains et 78 patients.

Deux tâches sont proposées sur ce corpus : d'une part le diagnostic de la maladie, qui vise à obtenir un bon score de classification binaire ; d'autre part une tâche d'estimation du *Mini-Mental State Examination* (MMSE), questionnaire évaluent les performances cognitives du sujet.

Les algorithmes proposés dans (Balagopalan *et coll.*, 2021; Haulcy et Glass, 2021; Meghanani *et coll.*, 2021b,a; Rohanian *et coll.*, 2021; Searle *et coll.*, 2020; Shah *et coll.*, 2021; Syed *et coll.*, 2020) permettent des RMSE inférieures à 6 avec le MMSE et des scores de classification correcte de plus de 80%.

ADReSSo D2* Introduit dans (Luz *et coll.*, 2021), ce corpus étends le sous-corpus du précédent et contient 83 patients et 83 sujets sains.

Les systèmes proposés sur ce corpus permettent des RMSE sur le MMSE inférieures à 4.3 (Syed *et coll.*, 2021c; Rohanian *et coll.*, 2021; Pappagari *et coll.*, 2021) et des performances de classification supérieures 80% (Chen *et coll.*, 2021; Pan *et coll.*, 2021; Pappagari *et coll.*, 2021; Rohanian *et coll.*, 2021; Syed *et coll.*, 2021c).

CCC La *Carolina Conversations Collection* est un large corpus de conversations en anglais avec une population âgée, dont une cohorte comprend des patients atteints de la maladie d'Alzheimer (Pope et Davis, 2011).

Lors de son exploitation pour créer des systèmes d'apprentissage pour la détection de la maladie à partir de la voix des sujets, le corpus est réduit à 16 sujets sains et 30 patients.

Les trois études identifiées utilisant ce corpus atteignent des performances de classification supérieures à 80% (la Fuente Garcia *et coll.*, 2020; Nasreen *et coll.*, 2021b,a).

Une des limites de ce corpus repose sur le fait que les sujets dits « sains » ont une autre pathologie telle que des problèmes de coeur, du diabète, un cancer, de l'arthrose, une leucémie, etc., qui influe également sur la voix des sujets.

Corpus pour le pronostic

ADReSSo D1 Dans ce corpus, publié en même temps que le ADReSSo D2, une nouvelle tâche est introduite : celle de la progression de la maladie sur une durée de 2 ans. Ce corpus contient ainsi les enregistrements de 32 patients atteints de la maladie d'Alzheimer, dont on a mesuré le MMSE à deux ans d'intervalle. Ceux-ci sont alors séparés en deux classes : ceux ayant des performances cognitives ayant décliné (différence de MMSE > 5, n = 10) et ceux n'ayant pas décliné (n = 22).

Sur cette tâche de classification, les systèmes proposés par (Rohanian *et coll.*, 2021; Syed *et coll.*, 2021c) atteignent des performances de classification respectives de 67% et 73.80% (moyenne des F1-scores).

Corpus pour le diagnostic différentiel

Les principaux corpus proposant des tâches de diagnostic différentiel proposent la différenciation entre les sujets ayant un diagnostic de la maladie d'Alzheimer établis et ceux présentant son principal symptôme, la déficience cognitive légère (*Mild Cognitive Impairement* – MCI).

C'est par exemple le cas des corpus introduits par Clarke *et coll.* (2021); De Looze *et coll.* (2021); Jang *et coll.* (2021); Lindsay *et coll.* (2021); Sangchocanonta *et coll.* (2021); Yamada *et coll.* (2021a), qui contiennent tous les enregistrements de sujets sains, de patients atteints de la maladie d'Alzheimer et de sujets présentant un déficit cognitif, mais n'ayant pas été diagnostiqués comme ayant cette pathologie.

Une étude se différencie des précédentes, en proposant l'étude d'enregistrements de patients atteints de la maladie d'Alzheimer, de sujets atteints de troubles cognitifs légers, mais aussi d'une population atteinte de troubles cognitifs sévères non-Alzheimer (Kim *et coll.*, 2020a).

Autre corpus De nombreux autres corpus non détaillés ici ont été conçus pour la détection d'Alzheimer dans la voix, comme par exemple dans (Momeni et Rahmani, 2021; Nasrolahzadeh *et coll.*, 2020; Sadeghian *et coll.*, 2021; Yamada *et coll.*, 2021b).

2.6 Troubles psychiatriques

2.6.1 Hypothèse

Comme aperçu dans le chapitre précédent, les processus qui conduisent à la production de la parole ne se limitent pas à des procédés musculaires et neurolinguistiques, mais font également intervenir des processus cognitifs : « la voix est une fenêtre sur l'esprit » (Low *et coll.*, 2020).

2.6.2 Crise suicidaire

Le premier trouble psychique que nous décrivons ici est la crise suicidaire, avec ou sans tentative de suicide (TS). Une revue des systèmes proposés pour cette tâche de classification est proposée dans (Ji *et coll.*, 2021).

Pensées suicidaires avec TS ou historique de TS

Nous avons identifié trois corpus contenant des enregistrements de sujets ayant des pensées suicidaires ou un historique de TS.

Le corpus proposé dans (Ries *et coll.*, 2021) contient les enregistrements d'entretiens de 20 patients admis aux urgences psychiatriques après une tentative de suicide, et dont la gravité des pensées suicidaires est annotée avec l'Échelle d'idéation suicidaire de Beck (Dozois et Covin, 2004).

Dans (Stasak *et coll.*, 2021a), 20 sujets sains, 74 patients avec pensées suicidaires [tels que mesurés par l'échelle C-SSRS (Posner *et coll.*, 2008)], et 152 sujets avec un historique de TS

ont été inclus. Ces sujets ont été enregistrés sur une tâche de lecture de 21 phrases en anglais conçues pour induire un large spectre d'émotions chez les sujets, tels que « je veux vivre » ou « je veux mourir ». L'algorithme proposé permet de détecter les personnes avec des pensées suicidaires avec des performances de 73% et les personnes avec un historique de TS avec une performance 78%.

Dans (Zhang *et coll.*, 2020a), 222 sujets ont été enregistrés avec des smartphones sur la lecture de la phrase "The quick brown fox jumps over the lazy dog" et sur 30s de parole libre. Le système proposé permet la détection des idées suicidaire, telles que mesurées par l'item 9 de la *Patient Health Questionnaire* (PHQ), avec une AUC de 0.82.

Vétérans de guerre

Une population particulière pour cette tâche est les vétérans de guerre, qui sont particulièrement à risque. Dans l'étude proposée par Belouali *et coll.* (2021), une tablette a été proposée à une population de vétérans étatsuniens de la guerre du Golfe afin de suivre différents paramètres, dont l'idéation suicidaire (telle que mesurée par l'item 9 de la PHQ-9). Ils ont été enregistrés en train de répondre à des questionnaires à voix haute sur la tablette (n = 588 échantillons, 88 échantillons correspondants à un item 9 de la PHQ-9 autre que "pas du tout"). Le système de classification proposé permet d'atteindre une AUC de 80%.

Étudiants

Une autre population particulièrement à risque est les étudiants.

Dans le corpus étatsunien proposé par Cohen *et coll.* (2020), 60 étudiants provenant de 8 écoles différentes ont été enregistrés lors d'une séance de psychothérapie, durant laquelle leur niveau de risque suicidaire a été évalué. Sur les 60 étudiants inclus, 29 présentent des idées suicidaires (items 9 et 12 de la PHQ-A). Le système proposé permet d'atteindre une AUC de 0.78.

Dans (Figueroa Saavedra *et coll.*, 2021), ce sont 60 étudiants chiliens de 18 et 19 ans qui ont été enregistrés en train de vocaliser la voyelle /a/, de lire à voix haute un texte (*Grandfather passage*) et de répondre aux questions « Quel est le meilleur/pire qui vous soit arrivé récemment ? ». La méthode statistique proposée dans l'article permet de différencier les 48 étudiants considérés comme à risque (tel que mesuré par un score supérieur ou égal à 5 à l'échelle de risque suicidaire d'Okasha) des autres.

2.6.3 Depression

La tâche la plus représentée pour la détection de troubles psychiatriques dans la voix est la détection du trouble dépressif.

Corpus en conditions contrôlées

DAIC-WOZ Le corpus le plus utilisé pour la détection de la dépression dans la voix – et même toutes pathologies confondues – est le *Distress Analysis Interview Corpus of Human and Computer Interviews* (DAIC). Ce corpus est divisé en plusieurs sous-corpus, suivant les interactions proposées aux sujets : entretien avec un psychiatre, avec un agent conversationnel complètement automatique, ou avec un agent virtuel contrôlé à distance (*Wizard of Oz* – WOZ) (Gratch *et coll.*, 2014). Dans ce dernier sous-corpus, appelé *DAIC-WOZ*, l'agent est

animé par deux techniciens, qui contrôlent respectivement les questions posées par l'agent, et les gestes de celui-ci.

Ce sous-corpus contient les enregistrements de 189 vétérans de guerre étatsuniens, enregistrés pendant leur interaction avec le WOZ. Leur niveau de dépression est évalué grâce à la PHQ-8, pour laquelle un score supérieur ou égal à 10 indique une forte probabilité d'épisode dépressif majeur (133 sujets sont donc catégorisés comme « dépressifs » et 56 comme « non-dépressifs »).

La richesse des systèmes proposés sur ce corpus permet d'atteindre des scores de F1 supérieurs à 80 % (Demiroglu *et coll.*, 2020; Solieman et Pustozerov, 2021; Srimadhur et Lalitha, 2020), des scores de classification correcte supérieurs à 70% avec une moyenne supérieure à 80% (Othmani *et coll.*, 2021; Aharonson *et coll.*, 2020; Deshpande *et coll.*, 2021; Huang *et coll.*, 2020b,c; Vázquez-Romero et Gallardo-Antolín, 2020; Wang *et coll.*, 2021b; Zhao *et coll.*, 2021) et des RMSE avec le PHQ-8 inférieurs à 6 (Aharonson *et coll.*, 2020; Yang *et coll.*, 2020; Zhang *et coll.*, 2020b; Zhao *et coll.*, 2020a,b).

AVEC 2013 Le deuxième corpus le plus utilisé pour cette tâche est le corpus proposé lors de la compétition AVEC 2013 (Valstar *et coll.*, 2013). Ce corpus contient les enregistrements de 292 sujets allemands issus de la population générale, dont le niveau de gravité de symptômes dépressifs est annoté avec le BDI-II, pour lequel un score supérieur ou égal à 18 est considéré comme pathologique. Lors de leur interaction avec un ordinateur, ils sont enregistrés sur des tâches très diversifiées : voyelles soutenues, comptage, lecture de textes (*Homo Faber* et *La Bise et le Soleil*, en allemand), des tâches de parole spontanée (meilleur cadeau reçu pendant l'enfance, événement triste pendant l'enfance, raconter une histoire imaginée), description d'image (*Cookie Theft*) ou encore, originalité de ce corpus, de chanter une chanson.

Les systèmes que nous avons identifiés et utilisant ce corpus atteignent 89.30% de classification correcte dans (Wang *et coll.*, 2021c) et des RMSE de 8.16 et 9.57 respectivement dans (Cummins *et coll.*, 2020) et (Zhao *et coll.*, 2020c).

Autres corpus De nombreux autres corpus en conditions contrôlées ont été identifiés, mais ne sont pas décrits ici. Pour plus d'informations, nous redirigeons le lecteur vers les articles suivants : (Cai *et coll.*, 2020; Aloshban *et coll.*, 2021; Di *et coll.*, 2021; Demiroglu *et coll.*, 2020; Harati *et coll.*, 2021; Uher *et coll.*, 2014; Higuchi *et coll.*, 2021; Kiss et Vicsi, 2017; Liu *et coll.*, 2020; Shin *et coll.*, 2021; Shinohara *et coll.*, 2020, 2021; Tao *et coll.*, 2020; Vitale *et coll.*, 2021; Stasak *et coll.*, 2021c,b)

Corpus en conditions écologiques

Cette partie présente quelques corpus collectés en conditions écologiques pour la tâche d'estimation de la dépression dans la voix. Dans tous les corpus de ce paragraphe (et sauf mention contraire), les enregistrements ont été effectués avec un smartphone, directement par le sujet, dans le lieu de son choix.

Sonde Health 2 Dans (Huang *et coll.*, 2018), les échantillons vocaux de 887 locuteurs tirés de la population générale étatsunienne ont été regroupés. Ceux-ci ont été enregistrés sur des voyelles soutenues, de la lecture de phrases, /pa/-/ta/-/ka/, la lecture d'un texte (*Rainbow passage*), et des tâches de parole spontanées demandant des charges cognitives différentes. À chaque enregistrement est associé un score de PHQ-9, pour lequel un score supérieur ou égal à 10 est considéré comme pathologique.

Les systèmes proposés sur ce corpus permettent des scores de classification de 72.7% dans (Huang *et coll.*, 2020b), un UAR de 68% dans (Huang *et coll.*, 2020c), et une RMSE normalisée de 0.58 dans (Stasak *et coll.*, 2021c).

Moodable Ce corpus, introduit dans (Dogrucu *et coll.*, 2020), contient les enregistrements de 266 locuteurs issus de la population générale étatsunienne, en train de lire la phrase "The quick brown fox jumps over the lazy dog" et annotés avec le PHQ-9. L'étude propose des systèmes de classification pour différentes valeurs de seuils pathologiques, avec des performances autour de 60%.

Little et al. 2021 Enfin, le corpus proposé dans (Little *et coll.*, 2021) se démarque des autres corpus enregistrés en conditions écologiques par le mode de collecte des échantillons vocaux. En effet, au lieu d'utiliser un smartphone et une application dédiée, les sujets ont porté pendant 7 jours consécutifs un bracelet enregistrant en continu un signal de faible qualité (mono, 8kHz).

Le corpus ainsi établi contient les enregistrements de 28 sujets sains et de 28 sujets atteints de dépression majeure (MINI DSM-IV), tous ayant plus de 60 ans et ayant été diagnostiqués pour une démence ou des troubles cognitifs : le mode opératoire semble particulièrement adapté à la population ciblée dans l'étude.

L'étude permet d'atteindre des scores de 100% de classification correcte avec des marqueurs tels que le nombre et la durée des interactions par jour.

Autres corpus Nous avons identifié trois autres corpus ayant été enregistrés en conditions écologiques : nous redirigeons le lecteur vers (Mundt *et coll.*, 2007, 2012; Zhang *et coll.*, 2020a) pour plus de détails.

Corpus pour le pronostic

Au-delà du diagnostic et de la différenciation binaire entre sujets dépressifs et non-dépressifs, une autre tâche proposée est l'estimation automatique du pronostic grâce à des marqueurs vocaux.

Hansen et col. 2022 C'est par exemple ce que proposent Hansen *et coll.* (2022), dans un corpus danois contenant les enregistrements de 40 sujets sains, de 42 patients en début d'épisode dépressif majeur (HRSD$>$ 17, non traités) et de 25 sujets en rémission. La tâche proposée est une tâche de narration de la vie personnelle du sujet. Les performances de classification du système proposé (AUC) sont de l'ordre de 0.71.

Kwon et col. 2021 Le corpus proposé dans (Kwon et Kim, 2021) contient les enregistrements de 76 patients étatsuniens diagnostiqués par des experts et évalués avec la MADRS. Ils sont enregistrés avec un smartphone en train de lire des mots et des phrases isolées, en train de compter, de réciter les mois de l'année, sur /pa/-/ta/-/ka/ ou encore la lecture d'un texte (*Grandfather passage*). Le système proposé permet de différencier les patients atteints de formes modérées de la pathologie et ceux atteints d'une forme plus lourde avec une performance de 78%.

Corpus pour le diagnostic différentiel

Enfin, une tâche particulièrement importante dans le cadre des troubles psychiatriques est le diagnostic différentiel entre les patients atteints de dépression et d'autres atteints de pathologies qui pourraient se confondre ou s'exprimer de la même manière dans la voix.

PROMPT Dans le corpus japonais proposé par Sumali *et coll.* (2020), 77 patients atteints de dépression (échelle de dépression de Hamilton (HAMD) > 8) et 43 patients atteints de démence sont enregistrés. Le système proposé permet de différencier les deux populations avec un score de 83.50%.

Klangporkun et col. 2021 Le corpus proposé par (Klangpornkun *et coll.*, 2021) contient les enregistrements de 27 sujets thaïlandais sains, de 27 patients atteints de diverses formes de dépression (dépression caractérisée, trouble bipolaire ou unipolaire en phase dépressive, mesurées avec le PHQ-9 et l'HAMD) et de 12 patients atteints d'autres pathologies décrites dans le DSM-IV ou le DMS-V (non précisées). Les sujets sont enregistrés en train de lire des textes, de compter, sur des tâches de parole spontanée (revue de symptômes, narration de soi) ou de description d'images, avec un smartphone en conditions écologiques. Le système proposé permet de distinguer les patients atteints de dépression des autres troubles mentaux avec une AUC de 0.94.

DementiaBank Enfin, la tâche proposée par Abdallah-Qasaimeh et Ratté (2021) n'est pas proprement parler du diagnostic différentiel, mais a retenu notre attention par le sujet, commun avec d'autres études citées précédemment, de la détection d'une pathologie dans une population âgée et comorbide. Ainsi, cette étude a pour but de classifier les enregistrements proposés dans la DementiaBank (Becker *et coll.*, 1994) précédemment présentée (cf. maladie d'Alzheimer), qui ne contient que des enregistrements de personnes âgées. L'étude classifie les sujets atteints de démence de ceux atteints de démence et de dépression (HAMD ≥ 8) avec une performance de 89.10%.

2.6.4 Anxiété

Alors que la détection de la dépression mobilise une grande partie des efforts de recherche de détection de pathologies dans la parole, nous n'avons trouvé que quatre études sur l'anxiété.

Düsseldorf Anxiety Corpus Le Düsseldorf Anxiety Corpus contient les enregistrements de 252 sujets allemands (Baird *et coll.*, 2020) en train de vocaliser la voyelle /a/ dans 5 intonations différentes. Le système proposé sur ce corpus permet une corrélation de Spearman de $\rho = 0.592$ entre les valeurs estimées et vérités terrain du *Beck Anxiety Inventory* (BAI).

Albuquerque et col. 2021 Dans l'étude menée par Albuquerque *et coll.* (2021), 112 sujets portugais sont enregistrés en train de lire 28 mots dissyllabiques et sur la description d'une image (*Cookie Theft*). Les analyses statistiques proposées ont permis de relier des descripteurs issus du signal de parole à l'anxiété telle que mesurée par l'*Hospital Anxiety and Depression* (HAD).

Kim et col. 2020 Dans (Kim *et coll.*, 2020b), 193 sujets sont enregistrés sur des tâches de discours spontané, de lecture de 4 phrases et de lecture de deux paragraphes phonétiquement équilibrés. Le système proposé permet un score de concordance (CCC) de 0.46 entre l'estimation du *State-trait anxiety inventory* (STAI) et de la *Generalized anxiety disorder* (GAD7) avec les vérités terrain correspondantes.

Di Matteo et col. 2021 Le corpus proposé par (Di Matteo *et coll.*, 2021) a beaucoup de similarité avec l'étude menée sur la dépression de fin de vie menée dans (Little *et coll.*, 2021). En effet, celui-ci contient les enregistrements en semi-continu des smartphones de 86 participants anglophones tirés de la population générale : toutes les 5 minutes, le smartphone sur lequel une application dédiée a été au préalable installée, enregistre durant 15s. La procédure dure 2 semaines et les catégories lexicales (estimées automatiquement) des mots employés par les sujets ont été corrélées au score d'anxiété (*Liebowitz Social Anxiety Scale*).

2.6.5 Stress post-traumatique

Une autre tâche, liée à l'anxiété, est la détection du stress post-traumatique.

Trauma survenu durant l'enfance Dans (Monti *et coll.*, 2021), 48 sujets sont enregistrés en train de vocaliser la voyelle /a/. Parmi eux, 12 n'ont pas rapporté de trauma durant l'enfance, tandis que 36 en reportent au moins 1. En utilisant des méthodes statistiques, l'étude a relié des perturbations de la voix à l'historique de déni des sujets ainsi qu'à la gravité des symptômes associés, mesurée par le *Childhood Trauma Questionnaire* et le *State-Trait Anxiety Inventory*.

Stress post-traumatique après admission aux urgences Dans (Schultebraucks *et coll.*, 2020), 81 patients parlant anglais, espagnol ou chinois sont enregistrés 1 mois après leur passage aux urgences et interrogés à l'aide de cinq questions sur la raison de leur arrivée aux urgences, et leurs attentes dans la vie. L'algorithme proposé permet, à partir des enregistrements vocaux et du visage des patients, de classifier les patients ayant un *PTSD Checklist for DSM-IV criteria* (PCL-4) supérieur ou égal à 33.

2.6.6 Trouble schizophrène

Au-delà des troubles liés à l'anxiété et à la dépression, des travaux portant sur le lien entre troubles schizophrènes et la voix commencent à apparaître dans la littérature.

Tang et col. 2021 Pour la détection automatique des troubles schizophrènes, l'étude proposée dans (Tang *et coll.*, 2021) aux États-Unis a collecté les enregistrements de 11 contrôles sains et de 20 patients atteints de Schizophrénie (diagnostiqués sur la base du DSM-5 et de la *Scale for the Assessment of Thought, Language and Communication* (TLC)), enregistrés en train de parler d'eux et de se rappeler de souvenirs positifs ou neutres.

Tan et col. 2021 Dans l'étude proposée par Tan *et coll.* (2021), 46 sujets sains et 43 patients atteints de schizophrénie (DSM-IV, *Positive And Negative Syndrome Scale (PANSS)*, TLC) ont été enregistrés sur des tâches de parole spontanées en anglais (par ex. fonctionnement quotidien, prise de repas, ...) pendant au moins 10 minutes. Le système proposé permet des performances de classification binaire de 95% et une AUC de 0.97.

de Boer et col. 2021 Enfin, l'étude proposée par de Boer *et coll.* (2021) en Allemagne collecte à la fois le statut pathologique des sujets (sains ou atteints de schizophrénie (DSM-IV)), mais aussi les symptômes positifs et négatifs associés (PANSS). Ainsi, grâce aux enregistrements de 284 sujets (142 sujets sains, 142 patients atteints de schizophrénie) en train de parler sur des sujets neutres (5-30 min), le système proposé permet une classification correcte du statut pathologique avec des performances de 86.2%, et une différenciation entre les patients présentant des symptômes à dominante positive et ceux présentant des symptômes à dominante négative de 74.2%.

2.6.7 Troubles bipolaires et unipolaires

Nous avons identifié sept corpus portant sur la détection des troubles bipolaires et unipolaires à partir de marqueurs vocaux.

Enregistrements en conditions contrôlées

Audio-visual Bipolar Disorder (BD) corpus Ce corpus, introduit dans (Çiftçi *et coll.*, 2018), a été proposé pour le challenge AVEC 2018. Collecté en Turquie, il contient les enregistrements de 39 sujets sains et de 50 patients atteints de troubles bipolaires. Ceux-ci sont hospitalisés et enregistrés lors des jours 0, 3, 7, 14, et 28. Annotée avec l'échelle d'évaluation de la manie de Young (YMRS), la base de données est constituée de 120 échantillons de sujets sains, 88 échantillons en phase maniaque (YMRS $\geq$ 20), 82 échantillons en phase hypomaniaque (7 $<$YMRS$<$ 20) et de 62 en phase de rémission (YMRS $\leq$ 7). Les sujets ont été enregistrés en train d'expliquer les raisons de leur volonté de participer à l'étude, en train de décrire des souvenirs tristes et heureux, et en train de décrire deux peintures (*Depression* de Van Gogh et *Home Sweet home* de Dengel).

Sur ce corpus multimodal, la partie audio du système proposé par Ceccarelli et Mahmoud (2021) atteint 80% d'UAR sur la tâche de classification binaire différenciant le statut pathologique des sujets.

Approches écologiques

Farrús et col. 2021 Dans l'étude proposée par Farrús *et coll.* (2021), 11 utilisateurs espagnols et italiens dans différentes phases du trouble bipolaire enregistrent à domicile une vidéo à l'aide de leur smartphone qui explique leur journée (durée moyenne : 26s). Annoté avec l'HAMD et la YMRS, le système proposé dans l'étude permet des RMSE respectivement de 3.9 et 1.99.

Faurholt-Jepsen et col. 2021 Le corpus proposé par Faurholt-Jepsen *et coll.* (2021) combine à la fois une approche écologique et la possibilité de faire du diagnostic différentiel entre troubles bipolaires et trouble unipolaire. Enregistrés lors d'appels téléphoniques avec des professionnels de santé avec des smartphones, 121 patients atteints de troubles bipolaires, 38 sujets sains et 48 patients atteints de troubles unipolaires (ICD-10) ont été enregistrés au Danemark et aux Pays-Bas.

L'approche proposée dans (Faurholt-Jepsen *et coll.*, 2021) permet une AUC de 0.54 sur la tâche de diagnostic différentiel « trouble bipolaire » vs « trouble unipolaire », tandis que le système proposé par Faurholt-Jepsen *et coll.* (2022) permet la différentiation des patients atteints de troubles bipolaires avec une AUC de 0.76, et une différentiation des sujets atteints de troubles unipolaires et des sujets sains avec une AUC de 0.72.

Diagnostic différentiel

Automated Monitoring of Symptoms Severity Interview (AMoSS-I) dataset Dans ce corpus introduit dans (Wang *et coll.*, 2020), 12 sujets sains, 17 patients atteints de trouble de la personnalité borderline et 21 patients atteints de trouble bipolaire (DSM-IV, BIS-11, IPDE) ont été enregistrés en train de répondre à des questions à propos de l'application de suivi d'humeur et de leur expérience des différents dispositifs qui leur ont été proposés (en anglais). Cependant, une des faiblesses de ce corpus réside dans les conditions d'enregistrements : alors que 32 sujets ont été enregistrés dans une salle de réunion, 18 d'entre eux ont été enregistrés en basse qualité au téléphone (8kHz, 8 bits).

Sur ce corpus, le système proposé par (Wang *et coll.*, 2021a) atteint un score F1 de 73.8% sur le diagnostic du trouble bipolaire, 82.7% sur le diagnostic de trouble de la personnalité borderline et de 71.0% sur le diagnostic différentiel entre les deux pathologies.

Yamamoto et col. 2020 Dans ce corpus japonais, 76 sujets sains, 79 patients atteints de dépression et 68 patients atteints de trouble bipolaire (DSM-5, HAMD-17) sont enregistrés lors d'une discussion avec un psychiatre sur leur vie quotidienne (Yamamoto *et coll.*, 2020). Les analyses statistiques proposées dans l'étude permettent de relier la voix à chacun des deux troubles.

CHI-MEI Dans ce corpus enregistré à Taiwan, 45 sujets sont enregistrés avec une webcam en train de décrire des vidéos induisant des émotions (Huang *et coll.*, 2020a). Ils sont répartis entre 15 sujets sains, 15 patients atteints de troubles bipolaires et 15 patients atteints de troubles unipolaires. Le système proposé atteint un score de 73.3% de classification correcte sur le problème de classification à trois classes.

Gosztolya et col. 2020 Enfin le corpus hongrois proposé par Gosztolya *et coll.* (2020) propose les enregistrements de 14 patients atteints de troubles bipolaires et de 25 patients atteints de troubles schizophréniques (DSM-5) sur une tâche de parole spontanée : « Racontez-moi votre journée d'hier. » Le système proposé permet une classification correcte à 81.1%.

2.6.8 TDAH

Nous avons identifié une seule étude proposant la détection du Trouble du Déficit de l'Attention avec/sans Hyperactivité (TDAH) dans la voix (Etter *et coll.*, 2021). Dans ce corpus étatsunien, 22 sujets sains et 28 patients atteints de TDAH (diagnostiqués avec le *Conners' Adult ADHD Diagnostic Interview* et la *Conners' Adult ADHD Rating Scale*) sont enregistrés sur la prononciation de /pa/-/ta/-/ka/, la lecture d'un texte (*Grandfather passage*) et une tâche de parole spontanée (raconter ses vacances). Les études statistiques effectuées ont permis de trouver des corrélations entre les mesures vocales et les symptômes rapportés par les sujets.

2.6.9 Troubles autistiques

Toutes les études sur la détection de l'autisme se sont concentrées sur des populations pédiatriques, période de la vie durant laquelle le diagnostic est généralement fait.

Talkar et col. 2020 Dans l'étude proposée par Talkar *et coll.* (2020), 5 sujets contrôles et 5 sujets atteints de troubles du spectre autistique (critère du DSM-5) sont enregistrés en train

de lire un texte en langue anglaise (*The caterpillar*). Le système proposé dans l'étude permet une différenciation des deux groupes avec une AUC de 1.0.

Eni et col. 2020 Dans l'étude proposée dans (Eni *et coll.*, 2020), 6 sujets sains, 10 sujets avec suspicion de trouble autistique et 56 sujets diagnostiqués (DSM-5) sont enregistrés lors des tâches proposées pour le diagnostic (*Autistim Diagnostic Observation Schedule* (ADOS) en hébreu). Le système proposé permet une estimation du score à l'ADOS avec un RMSE de 4.65.

Mohanta et col. 2020 Dans l'étude proposée par (Mohanta *et coll.*, 2020), 13 sujets diagnostiqués et 20 sujets contrôles sont enregistrés en train de prononcer des mots isolés en anglais. Le système proposé atteint un score de classification correcte de 96.50%.

ADOS dataset Ce corpus, introduit dans (Kim *et coll.*, 2021), contient les enregistrements de 165 sujets avec suspicion de trouble autistique lors de leur passation de l'ADOS. À l'issue de l'examen, 86 sujets ont été diagnostiqués positifs et 79 négatifs.

Le système proposé dans l'étude permet d'atteindre un score de classification (F1 macro) de 85%.

Asgari et col. 2021 Enfin, le corpus proposé dans (Asgari *et coll.*, 2021) contient les enregistrements de 118 sujets étatsuniens en train de passer l'ADOS, dont 90 diagnostiqués positivement (DSM-5). Le système proposé permet un score de classification de 73.3%.

2.6.10 Troubles du comportement alimentaire

Contrairement à la revue de Low *et coll.* (2020), nous n'avons pas identifié de travaux récents sur la détection ou le suivi de troubles alimentaires à partir de marqueurs vocaux.

2.7 Altération du fonctionnement général

La dernière partie de cette revue comprend toutes les altérations de l'état général d'un sujet : l'emprise de substances, de certaines émotions, du stress, ou encore les problèmes de sommeil (les troubles alimentaires n'ayant pas été identifiés dans la section précédente) semblent interférer de manière caractéristique avec la production de parole.

2.7.1 Intoxications

Alcool

Alcohol Language Corpus – ALC L'*Alcohol Language Corpus* (ALC) contient les enregistrements de 162 sujets allemands enregistrés sur une trentaine de tâches : lecture de phrases, de chiffres, commandes vocales, réponse spontanée à 5 questions. Le paradigme expérimental consiste à enregistrer les sujets sobres (état contrôle), puis après la prise d'alcool (vin ou bière) jusqu'à ce que le locuteur atteigne un niveau d'ébriété avec lequel il est confortable. Ensuite, une prise de sang et une mesure d'alcoolémie dans l'air expiré étaient effectuées. Le taux d'alcool était compris entre 0,05% et 0,25%.

Le système proposé dans (Breton et Cantin-Savoie, 2020) atteint un score de classification de 53.6% tandis que celui proposé dans (Shenoi *et coll.*, 2020) atteint 80% sur la classe *alcoolisé* et 94.13% sur la classe *sobre*.

Cannabis

Shamei et col. 2021 Dans (Shamei *et coll.*, 2021), 8 sujets étatsuniens (4M, 4F) sont enregistrés en train de soutenir 7 voyelles, avant et après la prise de cannabis qu'ils ont eux-mêmes fourni et dont la dose n'est pas contrôlée.

Le système proposé permet des scores de classification de 68.9% sur les sous-corpus composé des enregistrements des sujets F et de 67.9% sur les sujets M.

Vogel et col. 2021 La tâche proposée par Vogel *et coll.* (2021) est légèrement différente de la précédente puisqu'il s'agit ici de détecter un historique d'utilisation de cannabis, plutôt que l'emprise à un instant fixé. Ainsi, 31 sujets ayant des antécédents d'utilisation de cannabis dans un cadre récréatif et 40 contrôles ont été enregistrés sur un monologue d'une minute, sur la voyelle soutenue /a/, sur la prononciation de /pa/-/ta/-/ka/, sur la récitation des jours de la semaine et sur la lecture d'un texte (*Grandfather passage*). Les analyses statistiques ont permis de lier marqueurs vocaux et historique d'utilisation de cannabis thérapeutique.

MDMA

Agurto et col. 2020 Dans l'étude publiée par Agurto *et coll.* (2020), 31 sujets sont enregistrés en train de décrire une personne importante de leur vie et/ou sur un monologue sur le sujet de leur choix, pendant 5 minutes. Les sujets sont enregistrés durant 4 sessions, dans lesquelles ils reçoivent soit une dose de MDMA (deux doses différentes), soit une dose d'ocytocine, soit un placebo.

Les marqueurs vocaux et le classifieur conçus dans l'étude permettent des scores de classification tous supérieurs à 80% sur toutes les quatre tâches de classification binaire investiguées (différence entre les deux doses de MDMA, MDMA1 vs contrôle, MDMA2 vs contrôle, ocytocine vs contrôle).

2.7.2 Stress et état émotionnel

États émotionnels

La détection des états émotionnels dans la voix constitue un champ de recherche à part ("voice AND speech AND emotion*" retourne 122 résultats dans Web of Science et 273 sur Scopus pour la seule année 2021), et n'est pas traitée dans cette partie.

Il faut cependant tenir compte de l'influence de l'état émotionnel sur la voix, notamment lors de la conception des corpus (cf. chapitre 9).

Stress

Nous avons trouvé deux études récentes étudiant l'influence du stress sur la voix.

Baird et col. 2021 Dans l'étude menée par Baird *et coll.* (2021), trois corpus, tous enregistrés en allemand, sont utilisés pour la détection du stress dans la voix :

— le FAU-Trier Social Stress Test (FAU-TSST) (Baird *et coll.*, 2019), contenant les enregistrements de 43 sujets ;
— le Regensburg-Trier Social Stress Test (Reg-TSST) (Stappen *et coll.*, 2021), contenant les enregistrements de 27 sujets ;
— et le Ulm-Trier Social Stress Test (Ulm-TSST) corpus, contenant les enregistrements de 69 sujets.

Dans les trois corpus, les sujets sont soumis au test de stress social de Trier, qui consiste en une simulation d'entretien d'embauche. Les sujets complètent ensuite une tâche de calcul mental. À différents stades de l'étude, des tests salivaires sont effectués sur les sujets afin de doser leur concentration en cortisol.

Le système proposé par l'étude permet une estimation du niveau de cortisol à partir d'indices vocaux avec une corrélation $\rho = 0.770$ 10 minutes après la fin du test et $\rho = 0.698$ 20 minutes après le test.

König et col. 2020 Dans le corpus proposé par König *et coll.* (2021), 89 soignants travaillant en France durant la pandémie de Covid-19 ont été enregistrés avec un smartphone sur trois questions ouvertes (neutre, positive et négative). Ils remplissent à la fois le *Motivation, stress, affect questionnaire* (MSA) et sont annotés avec la *Perceived Stress Questionnaire* (PSQ) (hétéro questionnaire). Le système proposé permet une MAE de 5.31 avec le « score total de stress » proposé dans l'étude.

2.7.3 Pathologies liées au sommeil

Nous avons trouvé cinq corpus portant sur la détection des troubles du sommeil à partir de marqueurs vocaux.

Troubles du sommeil paradoxal (RBD)

Par leur intérêt en tant que potentiel marqueur de l'arrivée de la maladie de Parkinson, les épisodes de parasomnies en sommeil paradoxal ont été étudiés dans des études portant sur la maladie de Parkinson (cf. section 2.5.2).

Insomnie

Nous n'avons pas trouvé d'étude récente permettant l'estimation ou la détection de l'insomnie dans la voix.

Syndrome d'Apnée Obstructive du Sommeil (SAOS)

Pang et col. 2020 Dans l'étude menée par Pang *et coll.* (2020), 66 sujets hongkongais avec différentes sévérités d'apnée obstructive du sommeil (mesurée par l'*apnea-hypopnea index* durant une polysomnographie) ont été enregistrés en train de vocaliser les voyelles /a/, /e/, /i/, /o/ et /u/. Sur les 4 classes proposées (31 sujets contrôles, 13 avec une faible sévérité, 10 avec une sévérité moyenne, 12 avec un SAOS sévère), le classifieur proposé atteint un score de classification correcte de 95.45%.

Ding et al. 2021 Dans le corpus en Chinois proposé par Ding *et coll.* (2021), 151 sujets sont enregistrés sur la vocalisation des phonèmes /a/, /o/, /e/, /i/, /u/, /ü/, /en/ et /eng/. Deux tâches de classification correspondant à deux niveaux de sévérité du SAOS sont proposées : le SAOS léger (AHI > 10), et le SAOS sévère (AHI > 30). Le système proposé permet une classification correcte de l'ordre de 78.8% pour les deux tâches de classification.

Wei et col. 2021 Dans (Wei *et coll.*, 2021), 75 patients avec un SAOS sévère (AHI> 30, diagnostic : polysomnographie et *Voice Handicap Index*) et 46 sujets contrôles (STOP-BANG < 3) sont enregistrés en train de vocaliser la voyelle /i/. Le système proposé permet de différencier les patients des sujets contrôles avec une AUC de l'ordre de 0.80.

Botelho et col. 2021 Le corpus proposé par Botelho *et coll.* (2021) se démarque des précédents par l'origine du corpus, consistant en 40 vlogs YouTube dans lesquels les personnes disent être atteintes de SAOS (n=18) ou non (n=22). La seule vérité terrain est ainsi la plainte du sujet dans la vidéo. À l'aide uniquement des marqueurs audio, le système proposé dans l'étude a une performance de classification binaire de l'ordre de 67.5%.

Qualité de sommeil

Dans (Kim *et coll.*, 2020b), 193 sujets sont enregistrés sur une tâche de discours spontané, de lecture de 4 phrases et de lecture de deux paragraphes phonétiquement équilibrés. Le système proposé permet un score de concordance (CCC) de 0.50 entre l'estimation de l'Index de Qualité du Sommeil de Pittsburgh (PSQI) et la vérité terrain.

2.7.4 Fatigue

Nous n'avons pu identifier qu'une seule étude portant sur la détection de la fatigue à partir de marqueurs vocaux. Dans l'étude proposée par Yan *et coll.* (2021), 796 échantillons provenant d'échanges entre des pilotes d'avion et leur tour de contrôle sont collectés. Ceux-ci sont pour moitié étiquetés comme provenant d'un pilote fatigué, mais la mesure utilisée n'est pas précisée (il est probable qu'il s'agisse uniquement d'une déclaration du pilote signalant qu'il est fatigué).

Les systèmes de classification binaires présentés dans (Yan *et coll.*, 2021) et (Shen et Wei, 2021) atteignent des performances respectives de 94.25% et 92.82%.

2.8 Discussion et conclusion

Notre revue a ainsi permis d'identifier plus de trente pathologies différentes, avec cependant de très grandes disparités sur le nombre d'études : la détection de la dépression, de la maladie de Parkinson, d'Alzheimer ou des pathologies liées à l'appareil phonatoire semblent concentrer la grande majorité des efforts de recherche dans ce domaine. Dans cette section, nous soulevons quelques points de discussion, qui seront discutés exhaustivement dans les chapitres 9 et 18.

2.8.1 Nombre d'enregistrements par sujet

La grande majorité des études citées précédemment ne procèdent qu'à un seul enregistrement par sujet, pour détecter des pathologies qui s'expriment le plus souvent sur de longues périodes de temps et dont on ne connaît pas le mécanisme.

Si le mécanisme de la maladie est connu – comme par exemple dans le cas de la maladie d'Huntington, il est possible de chercher des marqueurs vocaux qui mesurent de manière spécifique les caractéristiques vocales déjà observées par les cliniciens, et pour lesquels un seul enregistrement par sujet permet de lier les marqueurs ainsi conçus à la pathologie.

À l'inverse, si le mécanisme de la pathologie et de son influence sur la voix n'est pas connu, il paraît difficile – voire impossible – de distinguer dans la voix les *traits* du locuteur, qui sont

stationnaires sur de longues durées – de leur *état*. Par exemple, dans le cadre de la détection de la dépression à partir de marqueurs vocaux, il paraît difficile de différencier la maladie, elle-même caractérisée par des critères de durée (15 jours pour l'épisode de dépression majeure dans le DSM-5), d'une émotion triste passagère. Cette problématique, intimement liée à la construction des bases de données, est discutée dans le chapitre 9.

2.8.2 Tâches de classification proposées

De même, excepté dans les études sur l'appareil phonatoire, et quelques études isolées dans les troubles psychiatriques et les maladies neurodégénératives, la tâche majoritairement représentée est le diagnostic de la pathologie. Dans cette tâche, l'objectif est généralement de classifier une population constituée de patients diagnostiqués et de sujets contrôles sains, et/ou d'estimer le degré de gravité ou d'avancement de la pathologie. La pertinence d'une telle approche est discutée dans le chapitre 18.

Cependant, d'autres tâches, dans de nombreux cas utiles à la pratique clinique, mériteraient plus d'attention de la communauté : le diagnostic différentiel, et le pronostic.

Diagnostic différentiel Dans cette tâche, le but est de différencier deux pathologies qui sont proches. Si la différentiation entre la pathologie et un des symptômes est bien représentée pour les maladies d'Alzheimer (avec le MCI) et la maladie de Parkinson (avec les troubles de parasomnie en sommeil paradoxal), il y a peu d'études proposant des tâches de diagnostic différentiel dans les troubles mentaux.

Pourtant, des systèmes utilisant la spécificité de certains marqueurs vocaux pour différencier les deux troubles seraient d'une grande utilité clinique pour les pathologies dont les diagnostics différentiels sont difficiles (par exemple la différence entre les troubles bipolaires et la schizophrénie, introduite dans (Gosztolya *et coll.*, 2020) ou entre les troubles bipolaires et borderline, introduite dans (Wang *et coll.*, 2020)).

Pronostic Le but d'une tâche de pronostic est d'estimer, à partir d'informations sur le patient récoltées durant une première visite, l'évolution de la pathologie sur une certaine durée. C'est ce qui est proposé dans le corpus ADReSSo D1 par Luz *et coll.* (2021) : les patients sont enregistrés à deux ans d'intervalle, et la tâche proposée est d'estimer la variation du MMSE sur cet intervalle. Si l'entraînement de ces systèmes nécessite d'avoir les enregistrements à la fois à la première et à la deuxième visite, le système implémenté en conditions cliniques ne récolte que la voix à la première visite et, sur la base de cet enregistrement et éventuellement d'autres informations médicales, permettra d'adapter le traitement en fonction des estimations des perspectives d'évolution de la maladie. De même que pour le diagnostic différentiel, ces systèmes seraient très utiles dans la pratique clinique, et particulièrement dans les maladies neurodégénératives dont l'estimation de rapidité d'évolution de la maladie est un critère de traitement très important.

Diagnostic d'une unique pathologie De plus, la focalisation de la communauté autour du diagnostic d'une seule pathologie crée parfois des situations où les sujets sains ou contrôles inclus dans l'étude ne sont sains au regard de la pathologie étudiée, mais pas d'autres pathologies pouvant également affecter la voix. C'est par exemple le cas du corpus Carolina conversations collection [CCC, (Pope et Davis, 2011)], utilisé dans la détection de la maladie d'Alzheimer, pour lequel les sujets dits « sains » sont atteints de pathologies cardiaques, de

cancers, d'arthrose... Or, toutes ces pathologies sont susceptibles d'affecter la voix. La notion de « sujet sain » est discutée plus précisément dans le chapitre 9.

2.8.3 Robustesse du diagnostic

Suivant les tâches proposées, nous avons pu observer une grande diversité dans la mesure médicale choisie pour annoter les données. Alors que dans la pathologie affectant l'appareil phonatoire et les systèmes respiratoires et cardiaques (excepté la Covid), la collecte de données se fait dans le cadre d'un environnement médical produisant un diagnostic robuste, d'autres tâches pour lesquelles le diagnostic est plus difficile ou moins accessible ne bénéficient pas toutes d'une telle qualité d'annotation.

Un des exemples les plus marquants en est le Cambridge dataset (Laguarta *et coll.*, 2020), contenant plusieurs milliers d'enregistrements de toux et de voix de sujets avec pour but la détection de la Covid-19. Le diagnostic est proposé selon trois modalités différentes :

— un test (PCR) ;
— le diagnostic réalisé par un médecin sur la base des symptômes ;
— une déclaration du sujet lui-même.

Au moment où l'étude a été publiée, les tests n'étaient pas aussi accessibles qu'ils l'ont été par la suite, ce qui explique la très faible proportion de diagnostics réalisés par ce moyen, pourtant le seul fiable pour la détection du virus. Alors qu'une personne présentant des symptômes peut potentiellement être infectée, la fiabilité des sujets se déclarant négatifs, car n'ayant pas de symptômes est, au mieux, douteuse. De même, que dire de l'utilité d'un système automatique entraîné à estimer non pas la Covid-19, diagnostiquée de manière fiable, mais la déclaration du sujet lui-même ?

Le choix d'une mesure médicale pour annoter les données est discuté dans le chapitre 9, tandis que les conséquences de ce choix sont discutées dans le chapitre 18.

2.8.4 Enregistrements en conditions écologiques vs en conditions contrôlées

Toutes les tâches identifiées ne bénéficient pas du même degré d'implémentation en conditions écologiques : alors que pour la détection de la maladie d'Alzheimer, les corpus les plus utilisés contiennent des enregistrements en conditions contrôlées, des tâches comme la détection de la dépression utilisent des corpus contenant des enregistrements effectués en conditions écologiques, avec des smartphones.

Des approches mixtes, utilisant un smartphone en conditions contrôlées, comme il y en a dans de nombreuses tâches nous semble une bonne approche première approche pour commencer la translation de ces outils du monde clinique à des conditions écologiques. Le choix de l'outil et de l'environnement d'enregistrement est discuté en détail dans le chapitre 9.

2.8.5 Tâches proposées

Les tâches proposées varient d'une pathologie à l'autre. Alors que pour les maladies de Parkinson et d'Huntington, les voyelles soutenues sont favorisées, les patients atteints de la maladie d'Alzheimer sont majoritairement enregistrés sur des questions ouvertes qui impliquent souvent la mémoire. Les tâches les plus représentées toutes pathologies confondues sont les suivantes :

— Voyelles soutenues : généralement la voyelle /a/, parfois avec différents volumes ou intonations (« en souriant »). Cette tâche, qui est la plus simple possible, permet d'estimer le fonctionnement des fonctions motrices de la voix.

— /pa/-/ta/-/ka/ : cette tache de diadococinésie orale demande plus d'agilité dans l'exécution que la simple voyelle tenue, et permet donc d'évaluer les capacités motrices et neuromotrices du sujet.
— Lecture de texte : la lecture d'un texte fait intervenir des fonctions cognitives de plus haut niveau que les deux tâches précédentes. Cette approche bénéficie en revanche d'une uniformisation du contenu de tous les enregistrements. Les textes les plus rencontrés dans la littérature sont la fable *La bise et le soleil*, *Rainbow* et le *Grandfather Passage*.
— Description d'image : une tâche de parole spontanée très utilisée dans les corpus est la description d'une image, avec une utilisation prédominante de l'image *Cookie Theft* du *Boston Diagnostic Aphasia Examination*. Cela permet un bon compromis entre parole spontanée et uniformisation des enregistrements, puisque tous les participants décrivent la même image, mais avec des mots différents.
— entretien avec médecin ou chercheur : cette dernière tâche est la tâche de parole spontanée la plus ouverte avec des tâches très variées, allant de la narration de son meilleur souvenir d'enfance à un véritable entretien clinique, voire une psychothérapie.

Le choix de la tâche sur laquelle enregistrer le sujet est discuté en détail dans le chapitre 9.

2.8.6 Langues

Si notre revue a permis de mettre en avant la diversité des langues dans lesquelles sont faites les études de détection de pathologie dans la voix, la majorité des études se sont limitées à une seule langue dans leur base de données. Nous avons identifié deux études utilisant des corpus multilingues, étudiant toutes deux la maladie de Parkinson.

Dans (Rusz *et coll.*, 2021), le corpus proposé contient des enregistrements en tchèque, anglais, allemand, français et italien de patients atteints de la maladie de Parkinson et de RBD. L'étude rapporte une absence de différence significative entre les marqueurs vocaux calculés sur chacune des langues.

Par ailleurs, l'étude menée par Vásquez-Correa *et coll.* (2019) a conclu que l'entraînement par transfert d'apprentissage (*transfer learning*) d'une langue à une autre permettait une meilleure robustesse des systèmes de détection de la maladie de Parkinson : une approche multilingue semble permettre d'identifier les caractéristiques vocales de la pathologie, indépendamment de la langue.

2.8.7 Conclusion

En conclusion, de grandes disparités subsistent entre les différentes pathologies qui font l'objet de l'attention de la communauté de traitement du signal vocal. Alors que la très grande majorité des systèmes ont pour but le diagnostic de pathologies dont les annotations ne sont pas toujours robustes, quelques études éparses commencent à proposer des tâches de pronostic, de diagnostic différentiel, ou à traiter la problématique imposée par la dépendance aux langues.

L'initiative ColiveVoice lancée par le *Luxembourg Institute of Health* en septembre 2021 s'est donné pour objectif la collecte d'une grande base de données, multilingue, multisymptômes et multicohortes, avec pour but de combler les lacunes identifiées précédemment et contribuant ainsi à façonner le futur de la détection de pathologies par la voix.

Dans la suite de ce document, nous proposons de nous concentrer sur un symptôme qui est présent dans de nombreuses pathologies et un marqueur d'altération de la santé : la somnolence.

Conclusion de la partie

Mécanismes de la production vocale

Dans cette partie, nous avons décrit succinctement le fonctionnement moteur et neurolinguistique de la production vocale, à la fois sous le prisme d'un modèle source-filtre, expliquant la production acoustique de la voix ; et sous le prisme d'un modèle neurolinguistique, explicitant les différents niveaux de compréhension, de production et de contrôle de la génération du signal de parole. Nous avons ensuite formalisé l'hypothèse que la voix et la parole puissent être des mesures des altérations des capacités cognitives dues à des pathologies et étudié les implications théoriques de cette hypothèse.

Revue des bases de données pour la détection de pathologies dans la voix

Nous avons proposé une large revue de la littérature des différentes pathologies détectables dans la parole, et des corpus associés. En plus de la très grande diversité des troubles identifiés, cette revue nous a permis de soulever des premières questions relatives à la construction de corpus permettant cette détection : nombre d'enregistrements par sujet, tâches de classification proposée (formulation du problème clinique sous la forme de problème de classification), robustesse du diagnostic, choix entre enregistrements écologiques sur smartphones ou en conditions contrôlées, tâches pour la production vocale, et enfin importance de la langue. Ces points seront discutés précisément dans les chapitres 9 et 18.

Prochaine partie

Dans la prochaine partie, nous proposons d'étudier plus spécifiquement une altération d'état qui est à la fois la cause et la conséquence de très nombreuses pathologies : la somnolence [1].

1. Celle-ci a été volontairement exclue de la revue précédente, l'état de l'art de la détection de la somnolence dans la voix étant traité au chapitre 11

Bibliographie de la partie

Abdallah-Qasaimeh, B., et Ratté, S. (**2021**). "Detecting depression in Alzheimer's disease and MCI by speech analysis," Journal of Theoretical and Applied Information Technology **99**(5), 1162–1171.

Agurto, C., Cecchi, G. A., Norel, R., Ostrand, R., Kirkpatrick, M., Baggott, M. J., Wardle, M. C., Wit, H. d., et Bedi, G. (**2020**). "Detection of acute 3,4-methylenedioxymethamphetamine (MDMA) effects across protocols using automated natural language processing," Neuropsychopharmacology **45**(5), 823–832, doi: 10.1038/s41386-020-0620-4.

Aharonson, V., Nooy, A. d., Bulkin, S., et Sessel, G. (**2020**). "Automated Classification of Depression Severity Using Speech - A Comparison of Two Machine Learning Architectures," dans *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, Oldenburg, Germany, pp. 1–4, doi: 10.1109/ICHI48887.2020.9374335.

AL-Dhief, F. T., Latiff, N. M. A., Baki, M. M., Malik, N. N. N. A., Sabri, N., et Albadr, M. A. A. (**2021**). "Voice Pathology Detection Using Support Vector Machine Based on Different Number of Voice Signals," dans *2021 26th IEEE Asia-Pacific Conference on Communications (APCC)*, pp. 1–6, doi: 10.1109/APCC49754.2021.9609830.

Al Ismail, M., Deshmukh, S., et Singh, R. (**2021**). "Detection of Covid-19 Through the Analysis of Vocal Fold Oscillations," dans *ICASSP 2021*, pp. 1035–1039, doi: 10.1109/ICASSP39728.2021.9414201.

Albuquerque, L., Valente, A. R. S., Teixeira, A., Figueiredo, D., Sa-Couto, P., et Oliveira, C. (**2021**). "Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan," PLOS ONE **16**(4), e0248842, doi: 10.1371/journal.pone.0248842.

Aloshban, N., Esposito, A., et Vinciarelli, A. (**2021**). "What You Say or How You Say It? Depression Detection Through Joint Modeling of Linguistic and Acoustic Aspects of Speech," Cognitive Computation doi: 10.1007/s12559-020-09808-3.

Alves, M., Silva, G., Bispo, B. C., Dajer, M. E., et Rodrigues, P. M. (**2021**). "Voice Disorders Detection Through Multiband Cepstral Features of Sustained Vowel," Journal of Voice doi: 10.1016/j.jvoice.2021.01.018.

Amato, F., Borzì, L., Olmo, G., et Orozco-Arroyave, J. R. (**2021**). "An algorithm for Parkinson's disease speech classification based on isolated words analysis," Health Information Science and Systems **9**(1), 32, doi: 10.1007/s13755-021-00162-8.

Arias-Londoño, J. D., Godino-Llorente, J. I., Markaki, M., et Stylianou, Y. (**2011**). "On combining information from modulation spectra and mel-frequency cepstral coefficients for automatic detection of pathological voices," Logopedics Phoniatrics Vocology **36**(2), 60–69, doi: 10.3109/14015439.2010.528788.

Arora, S., Lo, C., Hu, M., et Tsanas, A. (**2021**). "Smartphone Speech Testing for Symptom Assessment in Rapid Eye Movement Sleep Behavior Disorder and Parkinson's Disease," IEEE Access **9**, 44813–44824, doi: 10.1109/ACCESS.2021.3057715.

Asgari, M., Chen, L., et Fombonne, E. (**2021**). "Quantifying Voice Characteristics for Detecting Autism," Frontiers in Psychology **12**, 665096, doi: 10.3389/fpsyg.2021.665096.

Asiaee, M., Vahedian-Azimi, A., Atashi, S. S., Keramatfar, A., et Nourbakhsh, M. (**2020**). "Voice Quality Evaluation in Patients With COVID-19 : An Acoustic Analysis," Journal of Voice S0892–1997(20)30368–4, doi: 10.1016/j.jvoice.2020.09.024.

Asmae, O., Abdelhadi, R., Bouchaib, C., Sara, S., et Tajeddine, K. (**2020**). "Parkinson's Disease Identification using KNN and ANN Algorithms based on Voice Disorder," dans *IRASET 2020*, doi: 10.1109/IRASET48871.2020.9092228.

B. T., B., Hee, H. I., Teoh, O. H., Lee, K. P., Kapoor, S., Herremans, D., et Chen, J.-M. (**2020**). "Asthmatic versus healthy child classification based on cough and vocalised /a :/ sounds," The Journal of the Acoustical Society of America **148**(3), EL253–EL259, doi: 10.1121/10.0001933.

Baird, A., Amiriparian, S., Cummins, N., Sturmbauer, S., Janson, J., Messner, E.-M., Baumeister, H., Rohleder, N., et Schuller, B. W. (**2019**). "Using Speech to Predict Sequentially Measured Cortisol Levels During a Trier Social Stress Test," dans *Interspeech 2019*, pp. 534–538, doi: 10.21437/Interspeech.2019-1352.

Baird, A., Cummins, N., Schnieder, S., Krajewski, J., et Schuller, B. W. (**2020**). "An Evaluation of the Effect of Anxiety on Speech — Computational Prediction of Anxiety from Sustained Vowels," dans *Interspeech 2020*, pp. 4951–4955, doi: 10.21437/Interspeech.2020-1801.

Baird, A., Triantafyllopoulos, A., Zänkert, S., Ottl, S., Christ, L., Stappen, L., Konzok, J., Sturmbauer, S., Meßner, E.-M., Kudielka, B. M., Rohleder, N., Baumeister, H., et Schuller, B. W. (**2021**). "An Evaluation of Speech-Based Recognition of Emotional and Physiological Markers of Stress," Frontiers in Computer Science **3**, 750284, doi: 10.3389/fcomp.2021.750284.

Balagopalan, A., Eyre, B., Robin, J., Rudzicz, F., et Novikova, J. (**2021**). "Comparing Pre-trained and Feature-Based Models for Prediction of Alzheimer's Disease Based on Speech," Frontiers in Aging Neuroscience **13**, doi: 10.3389/fnagi.2021.635945.

Bartl-Pokorny, K. D., Pokorny, F. B., Batliner, A., Amiriparian, S., Semertzidou, A., Eyben, F., Kramer, E., Schmidt, F., Schönweiler, R., Wehler, M., et Schuller, B. W. (**2021**). "The voice of COVID-19 : Acoustic correlates of infection in sustained vowels," The Journal of the Acoustical Society of America **149**(6), 4377–4383, doi: 10.1121/10.0005194.

Becker, J. T., Boiler, F., Lopez, O. L., Saxton, J., et McGonigle, K. L. (**1994**). "The Natural History of Alzheimer's Disease : Description of Study Cohort and Accuracy of Diagnosis," Archives of Neurology **51**(6), 585–594, doi: 10.1001/archneur.1994.00540180063015.

Belouali, A., Gupta, S., Sourirajan, V., Yu, J., Allen, N., Alaoui, A., Dutton, M. A., et Reinhard, M. J. (**2021**). "Acoustic and language analysis of speech for suicidal ideation among US veterans," BioData Mining **14**(1), 11, doi: 10.1186/s13040-021-00245-y.

Boeve, B. F. (**2013**). "Idiopathic REM sleep behaviour disorder in the development of Parkinson's disease," The Lancet Neurology **12**(5), 469–482, doi: 10.1016/S1474-4422(13)70054-1.

Bot, B. M., Suver, C., Neto, E. C., Kellen, M., Klein, A., Bare, C., Doerr, M., Pratap, A., Wilbanks, J., Dorsey, E. R., Friend, S. H., et Trister, A. D. (**2016**). "The mPower study, Parkinson disease mobile data collected using ResearchKit," Scientific Data **3**(1), 160011, doi: 10.1038/sdata.2016.11.

Botelho, C., Abad, A., Schultz, T., et Trancoso, I. (**2021**). "Visual Speech for Obstructive Sleep Apnea Detection," dans *Interspeech 2021*, pp. 2516–2520, doi: 10.21437/Interspeech.2021-1717.

Bourouhou, A., Jilbab, A., Nacir, C., et Hammouch, A. (**2021**). "Classification of cardiovascular diseases using dysphonia measurement in speech," Diagnostyka **Vol. 22, No. 1**, doi: 10.29354/diag/132586.

Breton, A., et Cantin-Savoie, G. (**2020**). "Automatisation de la détection de l'intoxication à l'alcool dans la parole," dans *Actes du congrès annuel de l'Association canadienne de linguistique 2020.*, p. 15.

Cai, H., Gao, Y., Sun, S., Li, N., Tian, F., Xiao, H., Li, J., Yang, Z., Li, X., Zhao, Q., Liu, Z., Yao, Z., Yang, M., Peng, H., Zhu, J., Zhang, X., Gao, G., Zheng, F., Li, R., Guo, Z., Ma, R., Yang, J., Zhang, L., Hu, X., Li, Y., et Hu, B. (**2020**). "MODMA dataset : a Multi-modal Open Dataset for Mental-disorder Analysis," arXiv :2002.09283 [cs, q-bio] .

Carrón, J., Campos-Roca, Y., Madruga, M., et Pérez, C. J. (**2021**). "A mobile-assisted voice condition analysis system for Parkinson's disease : assessment of usability conditions," Biomedical Engineering Online **20**(1), 114, doi: 10.1186/s12938-021-00951-y.

Ceccarelli, F., et Mahmoud, M. (**2021**). "Multimodal temporal machine learning for Bipolar Disorder and Depression Recognition," Pattern Analysis and Applications doi: 10.1007/s10044-021-01001-y.

Cesari, U., De Pietro, G., Marciano, E., Niri, C., Sannino, G., et Verde, L. (**2018**). "A new database of healthy and pathological voices," Computers & Electrical Engineering **68**, 310–321, doi: 10.1016/j.compeleceng.2018.04.008.

Chen, J., Ye, J., Tang, F., et Zhou, J. (**2021**). "Automatic detection of Alzheimer's disease using spontaneous speech only," dans *Interspeech 2021*, Vol. 6, pp. 4181–4185, doi: 10.21437/Interspeech.2021-2002.

Chiaramonte, R., et Vecchio, M. (**2021**). "A Systematic Review of Measures of Dysarthria Severity in Stroke Patients," PM&R **13**(3), 314–324, doi: 10.1002/pmrj.12469.

Chiba, T., et Kajiyama, M. (**1958**). *The vowel : Its nature and structure*, **652** (Phonetic society of Japan Tokyo).

Chui, K. T., Lytras, M. D., et Vasant, P. (**2020**). "Combined Generative Adversarial Network and Fuzzy C-Means Clustering for Multi-Class Voice Disorder Detection with an Imbalanced Dataset," Applied Sciences **10**(13), 4571, doi: 10.3390/app10134571.

Çiftçi, E., Kaya, H., Güleç, H., et Salah, A. A. (**2018**). "The turkish audio-visual bipolar disorder corpus," dans *2018 First Asian Conference on Affective Computing and Intelligent Interaction (ACII Asia)*, IEEE, pp. 1–6.

Clarke, N., Barrick, T., et Garrard, P. (**2021**). "A Comparison of Connected Speech Tasks for Detecting Early Alzheimer's Disease and Mild Cognitive Impairment Using Natural Language Processing and Machine Learning," Frontiers in Computer Science **3**, doi: 10.3389/fcomp.2021.634360.

Claverie, B. (**2021**). Cognition et Formation *Des théories pour la cognition. Différences et complémentarité des paradigmes*, l'harmattan éd.

Cohen, J., Wright-Berryman, J., Rohlfs, L., Wright, D., Campbell, M., Gingrich, D., Santel, D., et Pestian, J. (**2020**). "A Feasibility Study Using a Machine Learning Suicide Risk Prediction Model Based on Open-Ended Interview Language in Adolescent Therapy Sessions," International Journal of Environmental Research and Public Health **17**(21), 8187, doi: 10.3390/ijerph17218187.

Cummins, N., Sethu, V., Epps, J., Williamson, J. R., Quatieri, T. F., et Krajewski, J. (**2020**). "Generalized Two-Stage Rank Regression Framework for Depression Score Prediction from Speech," IEEE Transactions on Affective Computing **11**(2), 272–283, doi: 10.1109/TAFFC.2017.2766145.

Danilovaité, M. (**2021**). "Perceptually Motivated Feature set for Vocal Folds State Assessment," dans *2020 IEEE 8th Workshop on Advances in Information, Electronic and Electrical Engineering (AIEEE)*, pp. 1–4, doi: 10.1109/AIEEE51419.2021.9435765.

Dash, T. K., Mishra, S., Panda, G., et Satapathy, S. C. (**2021**). "Detection of COVID-19 from speech signal using bio-inspired based cepstral features," Pattern Recognition **117**, 107999, doi: 10.1016/j.patcog.2021.107999.

de Boer, J. N., Voppel, A. E., Brederoo, S. G., Schnack, H. G., Truong, K. P., Wijnen, F. N. K., et Sommer, I. E. C. (**2021**). "Acoustic speech markers for schizophrenia-spectrum disorders : a diagnostic and symptom-recognition tool," Psychological Medicine 1–11, doi: 10.1017/S0033291721002804.

De Cock, E., Oostra, K., Bliki, L., Volkaerts, A.-S., Hemelsoet, D., De Herdt, V., et Batens, K. (**2021**). "Dysarthria following acute ischemic stroke : Prospective evaluation of characteristics, type and severity," International Journal of Language and Communication Disorders **56**(3), 549–557, doi: 10.1111/1460-6984.12607.

De Looze, C., Dehsarvi, A., Crosby, L., Vourdanou, A., Coen, R. F., Lawlor, B. A., et Reilly, R. B. (**2021**). "Cognitive and Structural Correlates of Conversational Speech Timing in Mild Cognitive Impairment and Mild-to-Moderate Alzheimer's Disease : Relevance for Early Detection Approaches," Frontiers in Aging Neuroscience **13**, 637404, doi: 10.3389/fnagi.2021.637404.

Demiroglu, C., Besirli, A., Ozkanca, Y., et Çelik, S. (**2020**). "Depression level assessment from multi-lingual conversational speech data using acoustic and text features," EURASIP Journal on Audio, Speech, and Music Processing **2020**(1), 17, doi: 10.1186/s13636-020-00182-4.

Denes, P. B., et Pinson, E. N. (**1963**). *The Speech Chain : The Physics and Biology of Spoken Language*, bell telephone laboratories, éd.

Deshmukh, S., Al Ismail, M., et Singh, R. (**2021**). "Interpreting Glottal Flow Dynamics for Detecting Covid-19 From Voice," dans *ICASSP 2021*, pp. 1055–1059, doi: 10.1109/ICASSP39728.2021.9414530.

Deshpande, Y., Patel, S., Lendhe, M., Chavan, M., et Koshy, R. (**2021**). "Emotion and Depression Detection from Speech," dans *ICT Analysis and Applications*, édité par S. Fong, N. Dey, et A. Joshi, **154** (Springer Singapore, Singapore), pp. 257–265, doi: 10.1007/978-981-15-8354-4_27.

Despotovic, V., Skovranek, T., et Schommer, C. (**2020**). "Speech Based Estimation of Parkinson's Disease Using Gaussian Processes and Automatic Relevance Determination," Neurocomputing **401**, 173–181, doi: 10.1016/j.neucom.2020.03.058.

Di, Y., Wang, J., Li, W., et Zhu, T. (**2021**). "Using i-vectors from voice features to identify major depressive disorder," Journal of Affective Disorders **288**, 161–166, doi: 10.1016/j.jad.2021.04.004.

Di Matteo, D., Wang, W., Fotinos, K., Lokuge, S., Yu, J., Sternat, T., Katzman, M. A., et Rose, J. (**2021**). "Smartphone-Detected Ambient Speech and Self-Reported Measures of Anxiety and Depression : Exploratory Observational Study," JMIR formative research **5**(1), e22723, doi: 10.2196/22723.

Ding, Y., Wang, J., Gao, J., Fang, Q., Li, Y., Xu, W., Wu, J., et Han, D. (**2021**). "Severity evaluation of obstructive sleep apnea based on speech features," Sleep & Breathing = Schlaf & Atmung **25**(2), 787–795, doi: 10.1007/s11325-020-02168-0.

Dogrucu, A., Perucic, A., Isaro, A., Ball, D., Toto, E., Rundensteiner, E. A., Agu, E., Davis-Martin, R., et Boudreaux, E. (**2020**). "Moodable : On feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data," Smart Health **17**, 100118, doi: 10.1016/j.smhl.2020.100118.

Dong, L., Liu, C., Mao, C., Chu, S., Li, J., Huang, X., et Gao, J. (**2020**). "Linguistic Characteristics of Mandarin-Speaking Huntington's Disease Patients," Chinese Medical Sciences Journal **35**(3), 207–214, doi: 10.24920/003669.

Dozois, D. J., et Covin, R. (**2004**). "The Beck Depression Inventory-II (BDI-II), Beck Hopelessness Scale (BHS), and Beck Scale for Suicide Ideation (BSS).," .

Eni, M., Dinstein, I., Ilan, M., Menashe, I., Meiri, G., et Zigel, Y. (**2020**). "Estimating Autism Severity in Young Children From Speech Signals Using a Deep Neural Network," IEEE Access **8**, 139489–139500, doi: 10.1109/ACCESS.2020.3012532.

Etter, N. M., Cadely, F. A., Peters, M. G., Dahm, C. R., et Neely, K. A. (**2021**). "Speech motor control and orofacial point pressure sensation in adults with ADHD," Neuroscience Letters **744**, 135592, doi: 10.1016/j.neulet.2020.135592.

Fagherazzi, G., Fischer, A., Ismael, M., et Despotovic, V. (**2021**). "Voice for Health : The Use of Vocal Biomarkers from Research to Clinical Practice," Digital Biomarkers 78–88, doi: 10.1159/000515346.

Fant, G. (**1970**). *Acoustic theory of speech production* (de Gruyter).

Farrús, M., Codina-Filbà, J., et Escudero, J. (**2021**). "Acoustic and prosodic information for home monitoring of bipolar disorder," Health Informatics Journal **27**(1), 1460458220972755, doi: 10.1177/1460458220972755.

Faurholt-Jepsen, M., Rohani, D. A., Busk, J., Tønning, M. L., Vinberg, M., Bardram, J. E., et Kessing, L. V. (**2022**). "Discriminating between patients with unipolar disorder, bipolar disorder, and healthy control individuals based on voice features collected from naturalistic smartphone calls," Acta Psychiatrica Scandinavica **145**(3), 255–267, doi: 10.1111/acps.13391.

Faurholt-Jepsen, M., Rohani, D. A., Busk, J., Vinberg, M., Bardram, J. E., et Kessing, L. V. (**2021**). "Voice analyses using smartphone-based data in patients with bipolar disorder, unaffected relatives and healthy control individuals, and during different affective states," International Journal of Bipolar Disorders **9**(1), 38, doi: 10.1186/s40345-021-00243-3.

Fazeli, M., Moradi, N., Soltani, M., Naderifar, E., Majdinasab, N., Latifi, S. M., et Dastoorpour, M. (**2020**). "Dysphonia Characteristics and Vowel Impairment in Relation to Neurological Status in Patients with Multiple Sclerosis," Journal of Voice **34**(3), 364–370, doi: 10.1016/j.jvoice.2018.09.018.

Figueroa Saavedra, C., Otzen Hernández, T., Alarcón Godoy, C., Ríos Pérez, A., Frugone Salinas, D., et Lagos Hernández, R. (**2021**). "Association between suicidal ideation and acoustic parameters of university students' voice and speech : a pilot study," Logopedics Phoniatrics Vocology **46**(2), 55–62, doi: 10.1080/14015439.2020.1733075.

Gidaye, G., Nirmal, J., Ezzine, K., et Frikha, M. (**2020**). "Wavelet sub-band features for voice disorder detection and classification," Multimedia Tools and Applications **79**(39), 28499–28523, doi: 10.1007/s11042-020-09424-1.

Gosztolya, G., Bagi, A., Szalóki, S., Szendi, I., et Hoffmann, I. (**2020**). "Making a Distinction Between Schizophrenia and Bipolar Disorder Based on Temporal Parameters in Spontaneous Speech," dans *Interspeech 2020*, pp. 4566–4570, doi: 10.21437/Interspeech.2020-49.

Gour, G., Dr.V.Udayashankara, et Badak, D. (**2020**). "Voice-Disorder Identification of Laryngeal Cancer Patients," International Journal of Advanced Computer Science and Applications **11**, 352–358, doi: 10.14569/IJACSA.2020.0111145.

Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, S., et Morency, L.-P. (**2014**). "The Distress Analysis Interview Corpus of human and computer interviews," dans *LREC 2014*, Reykjavik, Iceland, pp. 3123–3128.

Haider, F., De La Fuente, S., et Luz, S. (**2020**). "An Assessment of Paralinguistic Acoustic Features for Detection of Alzheimer's Dementia in Spontaneous Speech," IEEE Journal on Selected Topics in Signal Processing **14**(2), 272–281, doi: 10.1109/JSTSP.2019.2955022.

Hammami, I., Salhi, L., et Labidi, S. (**2020**). "Voice Pathologies Classification and Detection Using EMD-DWT Analysis Based on Higher Order Statistic Features," IRBM **41**(3), 161–171, doi: 10.1016/j.irbm.2019.11.004.

Han, J., Brown, C., Chauhan, J., Grammenos, A., Hasthanasombat, A., Spathis, D., Xia, T., Cicuta, P., et Mascolo, C. (**2021**). "Exploring Automatic COVID-19 Diagnosis via Voice and Symptoms from Crowdsourced Data," dans *ICASSP 2021*, Toronto, ON, Canada, pp. 8328–8332, doi: 10.1109/ICASSP39728.2021.9414576.

Hansen, L., Zhang, Y.-P., Wolf, D., Sechidis, K., Ladegaard, N., et Fusaroli, R. (**2022**). "A generalizable speech emotion recognition model reveals depression and remission," Acta Psychiatrica Scandinavica **145**(2), 186–199, doi: 10.1111/acps.13388.

Harar, P., Galaz, Z., Alonso-Hernandez, J. B., Mekyska, J., Burget, R., et Smekal, Z. (**2020**). "Towards robust voice pathology detection," Neural Computing and Applications **32**(20), 15747–15757, doi: 10.1007/s00521-018-3464-7.

Harati, A., Shriberg, E., Rutowski, T., Chlebek, P., Lu, Y., et Oliveira, R. (**2021**). "Speech-Based Depression Prediction Using Encoder-Weight-Only Transfer Learning and a Large Corpus," dans *ICASSP 2021*, Toronto, ON, Canada, pp. 7273–7277, doi: 10.1109/ICASSP39728.2021.9414208.

Haulcy, R., et Glass, J. (**2021**). "Classifying Alzheimer's Disease Using Audio and Text-Based Representations of Speech," Frontiers in Psychology **11**, doi: 10.3389/fpsyg.2020.624137.

Hecker, P., Pokorny, F. B., Bartl-Pokorny, K. D., Reichel, U., Ren, Z., Hantke, S., Eyben, F., Schuller, D. M., Arnrich, B., et Schuller, B. W. (**2021**). "Speaking Corona? Human and Machine Recognition of COVID-19 from Voice," dans *Interspeech 2021*, pp. 1029–1033, doi: 10.21437/Interspeech.2021-1771.

Higuchi, M., Sonota, N., Nakamura, M., Miyazaki, K., Shinohara, S., Omiya, Y., Takano, T., Mitsuyoshi, S., et Tokuno, S. (**2021**). "Performance Evaluation of a Voice-Based Depression Assessment System Considering the Number and Type of Input Utterances," Sensors **22**(1), 67, doi: 10.3390/s22010067.

Hu, H.-C., Chang, S.-Y., Wang, C.-H., Li, K.-J., Cho, H.-Y., Chen, Y.-T., Lu, C.-J., Tsai, T.-P., et Lee, O. K.-S. (**2021**). "Deep Learning Application for Vocal Fold Disease Prediction Through Voice Recognition : Preliminary Development Study," Journal of Medical Internet Research **23**(6), e25247, doi: 10.2196/25247.

Huang, K.-Y., Wu, C.-H., Su, M.-H., et Kuo, Y.-T. (**2020**a). "Detecting Unipolar and Bipolar Depressive Disorders from Elicited Speech Responses Using Latent Affective Structure Model," IEEE Transactions on Affective Computing **11**(3), 393–404, doi: 10.1109/TAFFC.2018.2803178.

Huang, Z., Epps, J., Dale, J., et Chen, M. (**2018**). "Depression Detection from short Uterrances via Diverse Smartphones in Natural Environmental Conditions," dans *Interspeech 2018*.

Huang, Z., Epps, J., Joachim, D., et Sethu, V. (**2020**b). "Natural Language Processing Methods for Acoustic and Landmark Event-Based Features in Speech-Based Depression Detection," IEEE Journal of Selected Topics in Signal Processing **14**(2), 435–448, doi: 10.1109/JSTSP.2019.2949419.

Huang, Z., Epps, J., Joachim, D., Stasak, B., Williamson, J. R., et Quatieri, T. F. (**2020**c). "Domain Adaptation for Enhancing Speech-Based Depression Detection in Natural Environmental Conditions Using Dilated CNNs," dans *Interspeech 2020*, pp. 4561–4565, doi: 10.21437/Interspeech.2020-3135.

Jang, H., Soroski, T., Rizzo, M., Barral, O., Harisinghani, A., Newton-Mason, S., Granby, S., Stutz da Cunha Vasco, T., Lewis, C., Tutt, P., Carenini, G., Conati, C., et Field, T. (**2021**). "Classification of Alzheimer's Disease Leveraging Multi-task Machine Learning Analysis of Speech and Eye-Movement Data," Frontiers in Human Neuroscience **15**, doi: 10.3389/fnhum.2021.716670.

Jeancolas, L., Petrovska-Delacrétaz, D., Mangone, G., Benkelfat, B.-E., Corvol, J.-C., Vidailhet, M., Lehéricy, S., et Benali, H. (**2021**). "X-Vectors : New Quantitative Biomarkers for Early Parkinson's Disease Detection From Speech," Frontiers in Neuroinformatics **15**, 578369, doi: 10.3389/fninf.2021.578369.

Ji, S., Pan, S., Li, X., Cambria, E., Long, G., et Huang, Z. (**2021**). "Suicidal Ideation Detection : A Review of Machine Learning Methods and Applications," IEEE Transactions on Computational Social Systems **8**(1), 214–226, doi: 10.1109/TCSS.2020.3021467.

Kadiri, S. R., et Alku, P. (**2020**). "Analysis and Detection of Pathological Voice Using Glottal Source Features," IEEE Journal of Selected Topics in Signal Processing **14**(2), 367–379, doi: 10.1109/JSTSP.2019.2957988.

Karaman, O., Cakin, H., Alhudhaif, A., et Polat, K. (**2021**). "Robust automated Parkinson disease detection based on voice signals with transfer learning," Expert Systems with Applications **178**, 115013, doi: 10.1016/j.eswa.2021.115013.

Kim, M., Kim, H., et Lim, J. (**2020**a). "Classification of Diagnosis of Alzheimer's Disease Based on Convolutional Layers of VGG16 Model using Speech Data," dans *International Conference on ICT Convergence*, Vol. 2020-October, pp. 456–459, doi: 10.1109/ICTC49870.2020.9289477.

Kim, S., Kwon, N., O'Connell, H., Fisk, N., Ferguson, S., et Bartlett, M. (**2020**b). ""How are you ?" Estimation of anxiety, sleep quality, and mood using computational voice analysis," dans *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, Montreal, QC, Canada, pp. 5369–5373, doi: 10.1109/EMBC44109.2020.9175788.

Kim, Y.-K., Lahiri, R., Nasir, M., Kim, S. H., Bishop, S., Lord, C., et Narayanan, S. S. (**2021**). "Analyzing Short Term Dynamic Speech Features for Understanding Behavioral Traits of Children with Autism Spectrum Disorder," dans *Interspeech 2021*, pp. 2916–2920, doi: 10.21437/Interspeech.2021-2111.

Kiran Reddy, M., Helkkula, P., Madhu Keerthana, Y., Kaitue, K., Minkkinen, M., Tolppanen, H., Nieminen, T., et Alku, P. (**2021**). "The automatic detection of heart failure using speech signals," Computer Speech & Language **69**, 101205, doi: 10.1016/j.csl.2021.101205.

Kiss, G., et Vicsi, K. (**2017**). "Mono- and multi-lingual depression prediction based on speech processing," International Journal of Speech Technology **20**(4), 919–935, doi: 10.1007/s10772-017-9455-8.

Klangpornkun, N., Ruangritchai, M., Munthuli, A., Onsuwan, C., Jaisin, K., Pattanaseri, K., Lortrakul, J., Thanakulakkarachai, P., Anansiripinyo, T., Amornlaksananon, A., Laohawee, S., et Tantibundhit, C. (**2021**). "Classification of Depression and Other Psychiatric Conditions Using Speech Features Extracted from a Thai Psychiatric and Verbal Screening Test," dans *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, Mexico, pp. 651–656, doi: 10.1109/EMBC46164.2021.9629571.

König, A., Riviere, K., Linz, N., Lindsay, H., Elbaum, J., Fabre, R., Derreumaux, A., et Robert, P. (**2021**). "Measuring Stress in Health Professionals Over the Phone Using Automatic Speech Analysis During the COVID-19 Pandemic : Observational Pilot Study," Journal of Medical Internet Research **23**(4), e24191, doi: 10.2196/24191.

Kröger, B. J., Stille, C. M., Blouw, P., Bekolay, T., et Stewart, T. C. (**2020**). "Hierarchical Sequencing and Feedforward and Feedback Control Mechanisms in Speech Production : A Preliminary Approach for Modeling Normal and Disordered Speech," Frontiers in Computational Neuroscience **14**(573554), doi: 10.3389/fncom.2020.573554.

Kwon, N., et Kim, S. (**2021**). "Depression Severity Detection Using Read Speech with a Divide-and-Conquer Approach," dans *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, Mexico, pp. 633–637, doi: 10.1109/EMBC46164.2021.9629868.

la Fuente Garcia, S. d., Haider, F., et Luz, S. (**2020**). "Cross-corpus Feature Learning between Spontaneous Monologue and Dialogue for Automatic Classification of Alzheimer's Dementia Speech," dans *Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Annual International Conference*, Vol. 2020, pp. 5851–5855, doi: 10.1109/EMBC44109.2020.9176305.

Laganaro, M., Fougeron, C., Pernon, M., Levêque, N., Borel, S., Fournet, M., Catalano Chiuvé, S., Lopez, U., Trouville, R., Ménard, L., Burkhard, P. R., Assal, F., et Delvaux, V. (**2021**). "Sensitivity and specificity of an acoustic- and perceptual-based tool for assessing motor speech disorders in French : the MonPaGe-screening protocol," Clinical Linguistics & Phonetics **35**(11), 1060–1075, doi: 10.1080/02699206.2020.1865460.

Laguarta, J., Hueto, F., et Subirana, B. (**2020**). "COVID-19 Artificial Intelligence Diagnosis Using Only Cough Recordings," IEEE Open Journal of Engineering in Medicine and Biology **1**, 275–281, doi: 10.1109/OJEMB.2020.3026928.

Lévêque, N., Laganaro, M., Fougeron, C., Delvaux, V., Pernon, M., Borel, S., et Catalano, S. (**2016**). "MonPaGe : un protocole informatisé d'évaluation de la parole pathologique en langue française," Revue Neurologique **172**, A162–A163, doi: 10.1016/j.neurol.2016.01.386.

Lindsay, H., Tröger, J., et König, A. (**2021**). "Language Impairment in Alzheimer's Disease—Robust and Explainable Evidence for AD-Related Deterioration of Spontaneous Speech Through Multilingual Machine Learning," Frontiers in Aging Neuroscience **13**, 228, doi: 10.3389/fnagi.2021.642033.

Little, B., Alshabrawy, O., Stow, D., Ferrier, I. N., McNaney, R., Jackson, D. G., Ladha, K., Ladha, C., Ploetz, T., Bacardit, J., Olivier, P., Gallagher, P., et O'Brien, J. T. (**2021**). "Deep learning-based automated speech detection as a marker of social functioning in late-life depression," Psychological Medicine **51**(9), 1441–1450, doi: 10.1017/S0033291719003994.

Little, M., McSharry, P., Hunter, E., Spielman, J., et Ramig, L. (**2008**). "Suitability of dysphonia measurements for telemonitoring of Parkinson's disease," Nature Precedings 1–1, doi: 10.1038/npre.2008.2298.1.

Liu, Z., Guo, Z., Ling, Z., et Li, Y. (**2021**). "Detecting Alzheimer's disease from speech using neural networks with bottleneck features and data augmentation," dans *ICASSP 2021*, Vol. 2021-June, pp. 7323–7327, doi: 10.1109/ICASSP39728.2021.9413566.

Liu, Z., Xu, Y., Ding, Z., et Chen, Q. (**2020**). "Time-frequency Analysis Based on Hilbert-Huang Transform for Depression Recognition in Speech," dans *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, IEEE, Seoul, Korea (South), pp. 1072–1076, doi: 10.1109/BIBM49941.2020.9313587.

Low, D. M., Bentley, K. H., et Ghosh, S. S. (**2020**). "Automated assessment of psychiatric disorders using speech : A systematic review," Laryngoscope Investigative Otolaryngology **5**(1), 96–116, doi: 10.1002/lio2.354.

Luz, S., Haider, F., Fuente, S. d. l., Fromm, D., et MacWhinney, B. (**2020**). "Alzheimer's Dementia Recognition Through Spontaneous Speech : The ADReSS Challenge," dans *Interspeech 2020*, pp. 2172–2176, doi: 10.21437/Interspeech.2020-2571.

Luz, S., Haider, F., Fuente, S. d. l., Fromm, D., et MacWhinney, B. (**2021**). "Detecting cognitive decline using speech only : The ADReSSO Challenge," Rapport Technique, doi: 10.1101/2021.03.24.21254263.

Madruga, M., Campos-Roca, Y., et Pérez, C. J. (**2021**). "Multicondition Training for Noise-Robust Detection of Benign Vocal Fold Lesions From Recorded Speech," IEEE Access **9**, 1707–1722, doi: 10.1109/ACCESS.2020.3046873.

Mallela, J., Illa, A., Suhas, B., Udupa, S., Belur, Y., Atchayaram, N., Yadav, R., Reddy, P., Gope, D., et Ghosh, P. (**2020**). "Voice based classification of patients with Amyotrophic Lateral Sclerosis, Parkinson's Disease and Healthy Controls with CNN-LSTM using transfer learning," dans *ICASSP 2020*, Vol. 2020-May, pp. 6784–6788, doi: 10.1109/ICASSP40776.2020.9053682.

Meghanani, A., Anoop, C. S., et Ramakrishnan, A. G. (**2021**a). "Recognition of Alzheimer's Dementia From the Transcriptions of Spontaneous Speech Using fastText and CNN Models," Frontiers in Computer Science **3**, 7, doi: 10.3389/fcomp.2021.624558.

Meghanani, A., C. S., A., et Ramakrishnan, A. G. (**2021**b). "An Exploration of Log-Mel Spectrogram and MFCC Features for Alzheimer's Dementia Recognition from Spontaneous Speech," dans *2021 IEEE Spoken Language Technology Workshop (SLT)*, IEEE, Shenzhen, China, pp. 670–677, doi: 10.1109/SLT48900.2021.9383491.

Mendonça, J., Solera-Ureña, R., Abad, A., et Trancoso, I. (**2021**). "Using Self-Supervised Feature Extractors with Attention for Automatic COVID-19 Detection from Speech," arXiv :2107.00112 [eess.AS] doi: 10.48550/arXiv.2107.00112.

Mesallam, T. A., Farahat, M., Malki, K. H., Alsulaiman, M., Ali, Z., Al-nasheri, A., et Muhammad, G. (**2017**). "Development of the Arabic Voice Pathology Database and Its Evaluation by Using Speech Features and Machine Learning Algorithms," Journal of Healthcare Engineering **2017**, e8783751, doi: 10.1155/2017/8783751.

Min, S. N., Park, S. J., Im, J. N., et Subramaniyam, M. (**2020**). "A Bayesian Model for Prediction of Stroke with Voice Onset Time," IOP Conference Series : Materials Science and Engineering **912**(6), 062003, doi: 10.1088/1757-899X/912/6/062003.

Mohammed, M. A., Abdulkareem, K. H., Mostafa, S. A., Khanapi Abd Ghani, M., Maashi, M. S., Garcia-Zapirain, B., Oleagordia, I., Alhakami, H., et AL-Dhief, F. T. (**2020**). "Voice Pathology Detection and Classification Using Convolutional Neural Network Model," Applied Sciences **10**(11), 3723, doi: 10.3390/app10113723.

Mohanta, A., Mukherjee, P., et Mirtal, V. K. (**2020**). "Acoustic Features Characterization of Autism Speech for Automated Detection and Classification," dans *2020 National Conference on Communications (NCC)*, IEEE, Kharagpur, India, pp. 1–6, doi: 10.1109/NCC48643.2020.9056025.

Momeni, M., et Rahmani, M. (**2021**). "Speech signal analysis of alzheimer's diseases in farsi using auditory model system," Cognitive Neurodynamics **15**(3), 453–461, doi: 10.1007/s11571-020-09644-z.

Monti, E., D'Andrea, W., Freed, S., Kidd, D. C., Feuer, S., Carroll, L. M., et Castano, E. (**2021**). "Does Self-Reported Childhood Trauma Relate to Vocal Acoustic Measures ? Preliminary Findings at Trauma Recall," Journal of Nonverbal Behavior **45**(3), 389–408, doi: 10.1007/s10919-020-00355-x.

Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K., et Geralts, D. S. (**2007**). "Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology," Journal of Neurolinguistics **20**(1), 50–64, doi: 10.1016/j.jneuroling.2006.04.001.

Mundt, J. C., Vogel, A. P., Feltner, D. E., et Lenderking, W. R. (**2012**). "Vocal Acoustic Biomarkers of Depression Severity and Treatment Response," Biological Psychiatry **72**(7), 580–587, doi: 10.1016/j.biopsych.2012.03.015.

Narendra, N. P., et Alku, P. (**2020**). "Glottal Source Information for Pathological Voice Detection," IEEE Access **8**, 67745–67755, doi: 10.1109/ACCESS.2020.2986171.

Nasreen, S., Hough, J., et Purver, M. (**2021**a). "Detecting alzheimer's disease using interactional and acoustic features from spontaneous speech," dans *Interspeech 2021*, Vol. 1, pp. 306–310, doi: 10.21437/Interspeech.2021-1526.

Nasreen, S., Rohanian, M., Hough, J., et Purver, M. (**2021**b). "Alzheimer's Dementia Recognition From Spontaneous Speech Using Disfluency and Interactional Features," Frontiers in Computer Science **3**, 49, doi: 10.3389/fcomp.2021.640669.

Nasrolahzadeh, M., Haddadnia, J., et Rahnamayan, S. (**2020**). "Multi-Objective Optimization of Wavelet-Packet-Based Features in Pathological Diagnosis of Alzheimer Using Spontaneous Speech Signals," IEEE Access **8**, 112393–112406, doi: 10.1109/ACCESS.2020.3001426.

Noffs, G., Boonstra, F. M. C., Perera, T., Kolbe, S. C., Stankovich, J., Butzkueven, H., Evans, A., Vogel, A. P., et van der Walt, A. (**2020**). "Acoustic Speech Analytics Are Predictive of Cerebellar Dysfunction in Multiple Sclerosis," The Cerebellum **19**(5), 691–700, doi: 10.1007/s12311-020-01151-5.

Orozco-Arroyave, J. R., Arias-Londoño, J. D., Vargas-Bonilla, J. F., Gonzalez-Rátiva, M. C., et Nöth, E. (**2014**). "New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease," dans *LREC 2014*.

Othmani, A., Kadoch, D., Bentounes, K., Rejaibi, E., Alfred, R., et Hadid, A. (**2021**). "Towards Robust Deep Neural Networks for Affect and Depression Recognition from Speech," dans *Pattern Recognition. ICPR International Workshops and Challenges*, édité par A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, et R. Vezzani, **12662** (Springer International Publishing, Cham), pp. 5–19, doi: 10.1007/978-3-030-68790-8_1.

Pan, Y., Mirheidari, B., Harris, J., Thompson, J., Jones, M., Snowden, J., Blackburn, D., et Christensen, H. (**2021**). "Using the outputs of different automatic speech recognition paradigms for acoustic-and BERT-based Alzheimer's dementia detection through spontaneous speech," dans *Interspeech 2021*, Vol. 6, pp. 4216–4220, doi: 10.21437/Interspeech.2021-1519.

Pan, Y., Mirheidari, B., Reuber, M., Venneri, A., Blackburn, D., et Christensen, H. (**2020**). "Improving detection of Alzheimer's Disease using automatic speech recognition to identify high-quality segments for more robust feature extraction," dans *Interspeech 2020*, Vol. 2020-October, pp. 4961–4965, doi: 10.21437/Interspeech.2020-2698.

Pang, K.-G., Hsung, T.-C., Law, A. K.-W., et Choi, W. W. S. (**2020**). "Optimal vowels measurements for Obstructive Sleep Apnea Detection Using Speech Signals," dans *2020 IEEE 3rd International Conference on Information Communication and Signal Processing (ICICSP)*, pp. 143–147, doi: 10.1109/ICICSP50920.2020.9231972.

Pappagari, R., Cho, J., Joshi, S., Moro-Velazquez, L., Zelasko, P., Villalba, J., et Dehak, N. (**2021**). "Automatic detection and assessment of Alzheimer Disease using speech and language technologies in low-resource scenarios," dans *Interspeech 2021*, Vol. 6, pp. 4206–4210, doi: 10.21437/Interspeech.2021-1850.

Pinkas, G., Karny, Y., Malachi, A., Barkai, G., Bachar, G., et Aharonson, V. (**2020**). "SARS-CoV-2 Detection From Voice," IEEE Open Journal of Engineering in Medicine and Biology **1**, 268–274, doi: 10.1109/OJEMB.2020.3026468.

Pinyopodjanard, S., Suppakitjanusant, P., Lomprew, P., Kasemkosin, N., Chailurkit, L., et Ongphiphadhanakul, B. (**2021**). "Instrumental Acoustic Voice Characteristics in Adults with Type 2 Diabetes," Journal of Voice **35**(1), 116–121, doi: 10.1016/j.jvoice.2019.07.003.

Pope, C., et Davis, B. H. (**2011**). "Finding a balance : The Carolinas Conversation Collection," Corpus Linguistics and Linguistic Theory **7**(1), doi: 10.1515/cllt.2011.007.

Posner, K., Brent, D., Lucas, C., Gould, M., Stanley, B., Brown, G., Fisher, P., Zelazny, J., Burke, A., Oquendo, M., et others (**2008**). "Columbia-suicide severity rating scale (C-SSRS)," New York, NY : Columbia University Medical Center **10**.

Postuma, R. B., Gagnon, J. F., Rompre, S., et Montplaisir, J. Y. (**2010**). "Severity of REM atonia loss in idiopathic REM sleep behavior disorder predicts Parkinson disease," Neurology **74**(3), 239–244, doi: 10.1212/WNL.0b013e3181ca0166.

Postuma, R. B., Lang, A. E., Massicotte-Marquez, J., et Montplaisir, J. (**2006**). "Potential early markers of Parkinson disease in idiopathic REM sleep behavior disorder," Neurology **66**(6), 845–851, doi: 10.1212/01.wnl.0000203648.80727.5b.

Pulido, M. L. B., Hernández, J. B. A., Ballester, M. Á. F., González, C. M. T., Mekyska, J., et Smékal, Z. (**2020**). "Alzheimer's disease and automatic speech analysis : A review," Expert Systems with Applications **150**, 113213, doi: 10.1016/j.eswa.2020.113213.

Quatieri, T. F., Talkar, T., et Palmer, J. S. (**2020**). "A Framework for Biomarkers of COVID-19 Based on Coordination of Speech-Production Subsystems," IEEE Open Journal of Engineering in Medicine and Biology **1**, 203–206, doi: 10.1109/OJEMB.2020.2998051.

Ries, A., Abbas, A., Homan, S., Koesmahargyo, V., Yadav, V., Colla, M., Scheerer, H., Vetter, S., Seifritz, E., Scholz, U., Galatzer-Levy, I., et Kleim, B. (**2021**). "Validation of Visual and Auditory Digital Markers of Suicidality in Acutely Suicidal Psychiatric In-Patients," Biological Psychiatry **89**(9), S19–S20, doi: 10.1016/j.biopsych.2021.02.068.

Rizvi, D., Nissar, I., Masood, S., Ahmed, M., et Ahmad, F. (**2020**). "An LSTM based Deep learning model for voice-based detection of Parkinson's disease," International Journal of Advanced Science and Technology **29**, 337–343.

Rohanian, M., Hough, J., et Purver, M. (**2021**). "Alzheimer's dementia recognition using acoustic, lexical, disfluency and speech pause features robust to noisy inputs," dans *Interspeech 2021*, Vol. 6, pp. 4191–4195, doi: 10.21437/Interspeech.2021-1633.

Romana, A., Bandon, J., Carlozzi, N., Roberts, A., et Provost, E. M. (**2020**). "Classification of Manifest Huntington Disease using Vowel Distortion Measures," arXiv :2010.08503 [cs, eess] .

Roshanzamir, A., Aghajan, H., et Soleymani Baghshah, M. (**2021**). "Transformer-based deep neural network language models for Alzheimer's disease risk assessment from targeted speech," BMC Medical Informatics and Decision Making **21**(1), 92, doi: 10.1186/s12911-021-01456-3.

Rusz, J., Hlavnička, J., Novotný, M., Tykalová, T., Pelletier, A., Montplaisir, J., Gagnon, J.-F., Dušek, P., Galbiati, A., Marelli, S., Timm, P. C., Teigen, L. N., Janzen, A., Habibi, M., Stefani, A., Holzknecht, E., Seppi, K., Evangelista, E., Rassu, A. L., Dauvilliers, Y., Högl, B., Oertel, W., St. Louis, E. K., Ferini-Strambi, L., Růžička, E., Postuma, R. B., et Šonka, K. (**2021**). "Speech Biomarkers in Rapid Eye Movement Sleep Behavior Disorder and Parkinson Disease," Annals of Neurology **90**(1), 62–75, doi: 10.1002/ana.26085.

Sadeghian, R., Schaffer, J. D., et Zahorian, S. A. (**2021**). "Towards an Automatic Speech-Based Diagnostic Test for Alzheimer's Disease," Frontiers in Computer Science **3**, 13, doi: 10.3389/fcomp.2021.624594.

Sajal, M., Ehsan, M., Vaidyanathan, R., Wang, S., Aziz, T., et Mamun, K. (**2020**). "Telemonitoring Parkinson's disease using machine learning by combining tremor and voice analysis," Brain Informatics **7**(1), doi: 10.1186/s40708-020-00113-1.

Sakar, B. E., Isenkul, M. E., Sakar, C. O., Sertbas, A., Gurgen, F., Delil, S., Apaydin, H., et Kursun, O. (**2013**). "Collection and Analysis of a Parkinson Speech Dataset With Multiple Types of Sound Recordings," IEEE Journal of Biomedical and Health Informatics **17**(4), 828–834, doi: 10.1109/JBHI.2013.2245674.

Saleheen, N., Ahmed, T., Rahman, M. M., Nemati, E., Nathan, V., Vatanparvar, K., Blackstock, E., et Kuang, J. (**2020**). "Lung Function Estimation from a Monosyllabic Voice Segment Captured Using Smartphones," dans *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '20, pp. 1–11, doi: 10.1145/3379503.3403543.

Sangchocanonta, S., Vongsurakrai, S., Sroykhumpa, K., Ellermann, V., Munthuli, A., Anansiripinyo, T., Onsuwan, C., Hemrungrojn, S., Kosawat, K., et Tantibundhit, C. (**2021**). "Development of Thai Picture Description Task for Alzheimer's Screening using Part-of-Speech Tagging," dans *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, Mexico, pp. 2104–2109, doi: 10.1109/EMBC46164.2021.9629861.

Sara, J. D. S., Maor, E., Borlaug, B., Lewis, B. R., Orbelo, D., Lerman, L. O., et Lerman, A. (**2020**). "Non-invasive vocal biomarker is associated with pulmonary hypertension," PLOS ONE **15**(4), e0231441, doi: 10.1371/journal.pone.0231441.

Schuller, B., Steidl, S., Batliner, A., Hantke, S., Hönig, F., Orozco-Arroyave, J. R., Nöth, E., Zhang, Y., et Weninger, F. (**2015**). "The INTERSPEECH 2015 Computational Paralinguistics Challenge : Nativeness, Parkinson's & Eating Condition," dans *Interspeech 2015*, Dresden.

Schultebraucks, K., Yadav, V., Shalev, A. Y., Bonanno, G. A., et Galatzer-Levy, I. R. (**2020**). "Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood," Psychological Medicine 1–11, doi: 10.1017/S0033291720002718.

Schultz, B. G., Tarigoppula, V. S. A., Noffs, G., Rojas, S., van der Walt, A., Grayden, D. B., et Vogel, A. P. (**2021**). "Automatic speech recognition in neurodegenerative disease," International Journal of Speech Technology **24**(3), 771–779, doi: 10.1007/s10772-021-09836-w.

Searle, T., Ibrahim, Z., et Dobson, R. (**2020**). "Comparing natural language processing techniques for Alzheimer's dementia prediction in spontaneous speech," dans *Interspeech 2020*, Vol. 2020-October, pp. 2192–2196, doi: 10.21437/Interspeech.2020-2729.

Shah, Z., Sawalha, J., Tasnim, M., Qi, S.-a., Stroulia, E., et Greiner, R. (**2021**). "Learning Language and Acoustic Models for Identifying Alzheimer's Dementia From Speech," Frontiers in Computer Science **3**, 624–659, doi: 10.3389/fcomp.2021.624659.

Shamei, A., Sullivan, P. R., Liu, Y., Abdul-Mageed, M., et Gick, B. (**2021**). "Automated detection of cannabis intoxication from speech," Canadian Acoustics **49**(2).

Sharma, N., Krishnan, P., Kumar, R., Ramoji, S., Chetupalli, S. R., R., N., Ghosh, P. K., et Ganapathy, S. (**2020**). "Coswara – A Database of Breathing, Cough, and Voice Sounds for COVID-19 Diagnosis," Interspeech 2020 4811–4815, doi: 10.21437/Interspeech.2020-2768.

Shen, Z., et Wei, Y. (**2021**). "A high-precision feature extraction network of fatigue speech from air traffic controller radiotelephony based on improved deep learning," ICT Express **7**(4), 403–413, doi: 10.1016/j.icte.2021.01.002.

Shenoi, V. V., Kuchibhotla, S., et Kotturu, P. (**2020**). "An efficient state detection of a person by fusion of acoustic and alcoholic features using various classification algorithms," International Journal of Speech Technology **23**(3), 625–632, doi: 10.1007/s10772-020-09726-7.

Shimon, C., Shafat, G., Dangoor, I., et Ben-Shitrit, A. (**2021**). "Artificial intelligence enabled preliminary diagnosis for COVID-19 from voice cues and questionnaires," The Journal of the Acoustical Society of America **149**(2), 1120, doi: 10.1121/10.0003434.

Shin, H., Shivabasappa, P., et Koul, R. (**2021**). "Effect of clear speech intervention program on speech intelligibility in persons with idiopathic Parkinson's disease : A pilot study," International Journal of Speech-Language Pathology doi: 10.1080/17549507.2021.1943522.

Shinohara, S., Nakamura, M., Omiya, Y., Higuchi, M., Hagiwara, N., Mitsuyoshi, S., Toda, H., Saito, T., Tanichi, M., Yoshino, A., et Tokuno, S. (**2021**). "Depressive Mood Assessment Method Based on Emotion Level Derived from Voice : Comparison of Voice Features of Individuals with Major Depressive Disorders and Healthy Controls," International Journal of Environmental Research and Public Health **18**(10), 5435, doi: 10.3390/ijerph18105435.

Shinohara, S., Toda, H., Nakamura, M., Omiya, Y., Higuchi, M., Takano, T., Saito, T., Tanichi, M., Boku, S., Mitsuyoshi, S., So, M., Yoshino, A., et Tokuno, S. (**2020**). "Evaluation of the Severity of Major Depression Using a Voice Index for Emotional Arousal," Sensors **20**(18), 5041, doi: 10.3390/s20185041.

Solieman, H., et Pustozerov, E. A. (**2021**). "The Detection of Depression Using Multimodal Models Based on Text and Voice Quality Features," dans *2021 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (ElConRus)*, IEEE, St. Petersburg, Moscow, Russia, pp. 1843–1848, doi: 10.1109/ElConRus51938.2021.9396540.

Sonde One (**2021**). "Sonde Health" .

Sondhi, S., Salhan, A., Santoso, C. A., Doucoure, M., Dharmawan, D. M., Sureka, A., Natasha, B. N., Danusaputro, A. D., Dowson, N. S., Yap, M. S. L., Hadiwidjaja, M. A., Veeraraghavan, S. G., Hatta, A. Z. R., Lee, C., Megantara, R. A., Wihardja, A. N., Sharma, M., Lardizabal, E. L., Sondhi, L. J., Raina, R., Vashisth, S., et Hedwig, R. (**2021**). "Voice processing for COVID-19 scanning and prognostic indicator," Heliyon **7**(10), e08134, doi: 10.1016/j.heliyon.2021.e08134.

Srimadhur, N., et Lalitha, S. (**2020**). "An End-to-End Model for Detection and Assessment of Depression Levels using Speech," Procedia Computer Science **171**, 12–21, doi: 10.1016/j.procs.2020.04.003.

Stappen, L., Schumann, L., Sertolli, B., Baird, A., Weigell, B., Cambria, E., et Schuller, B. W. (**2021**). "Muse-toolbox : The multimodal sentiment analysis continuous annotation fusion and discrete class transformation toolbox," dans *Proceedings of the 2nd on Multimodal Sentiment Analysis Challenge*, pp. 75–82.

Stasak, B., Epps, J., Schatten, H. T., Miller, I. W., Provost, E. M., et Armey, M. F. (**2021**a). "Read speech voice quality and disfluency in individuals with recent suicidal ideation or suicide attempt," Speech Communication **132**, 10–20, doi: 10.1016/j.specom.2021.05.004.

Stasak, B., Huang, Z., Epps, J., et Dale, J. (**2021**b). "Depression Classification Using n-Gram Speech Errors from Manual and Automatic Stroop Color Test Transcripts," dans *43rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, doi: 10.1109/EMBC46164.2021.9629881.

Stasak, B., Huang, Z., Joachim, D., et Epps, J. (**2021**c). "Automatic Elicitation Compliance for Short-Duration Speech Based Depression Detection," dans *ICASSP 2021*, Toronto, ON, Canada, pp. 7283–7287, doi: 10.1109/ICASSP39728.2021.9414366.

Stasak, B., Huang, Z., Razavi, S., Joachim, D., et Epps, J. (**2021**d). "Automatic Detection of COVID-19 Based on Short-Duration Acoustic Smartphone Speech Analysis," Journal of Healthcare Informatics Research **5**(2), 201–217, doi: 10.1007/s41666-020-00090-4.

Sumali, B., Mitsukura, Y., Liang, K.-c., Yoshimura, M., Kitazawa, M., Takamiya, A., Fujita, T., Mimura, M., et Kishimoto, T. (**2020**). "Speech Quality Feature Analysis for Classification of Depression and Dementia Patients," Sensors **20**(12), 3599, doi: 10.3390/s20123599.

Suppakitjanusant, P., Sungkanuparph, S., Wongsinin, T., Virapongsiri, S., Kasemkosin, N., Chailurkit, L., et Ongphiphadhanakul, B. (**2021**). "Identifying individuals with recent COVID-19 through voice classification using deep learning," Scientific Reports **11**(1), 19149, doi: 10.1038/s41598-021-98742-x.

Syed, M. S. S., Syed, Z. S., Lech, M., et Pirogova, E. (**2020**). "Automated Screening for Alzheimer's Dementia Through Spontaneous Speech," dans *Interspeech 2020*, pp. 2222–2226, doi: 10.21437/Interspeech.2020-3158.

Syed, S. A., Rashid, M., Hussain, S., Imtiaz, A., Abid, H., et Zahid, H. (**2021**a). "Inter classifier comparison to detect voice pathologies," Mathematical Biosciences and Engineering **18**(3), 2258–2273, doi: 10.3934/mbe.2021114.

Syed, S. A., Rashid, M., Hussain, S., et Zahid, H. (**2021**b). "Comparative Analysis of CNN and RNN for Voice Pathology Detection," BioMed Research International **2021**, e6635964, doi: 10.1155/2021/6635964.

Syed, Z., Syed, M., Lech, M., et Pirogova, E. (**2021**c). "Tackling the ADRESSO challenge 2021 : The MUET-RMIT system for Alzheimer's dementia recognition from spontaneous speech," dans *Interspeech 2021*, Vol. 6, pp. 4231–4235, doi: 10.21437/Interspeech.2021-1572.

Talkar, T., Williamson, J. R., Hannon, D. J., Rao, H. M., Yuditskaya, S., Claypool, K. T., Sturim, D., Nowinski, L., Saro, H., Stamm, C., Mody, M., Mcdougle, C. J., et Quatieri, T. F. (**2020**). "Assessment of Speech and Fine Motor Coordination in Children With Autism Spectrum Disorder," IEEE Access **8**, 127535–127545, doi: 10.1109/ACCESS.2020.3007348.

Tan, E. J., Meyer, D., Neill, E., et Rossell, S. L. (**2021**). "Investigating the diagnostic utility of speech patterns in schizophrenia and their symptom associations," Schizophrenia Research **238**, 91–98, doi: 10.1016/j.schres.2021.10.003.

Tang, S. X., Kriz, R., Cho, S., Park, S. J., Harowitz, J., Gur, R. E., Bhati, M. T., Wolf, D. H., Sedoc, J., et Liberman, M. Y. (**2021**). "Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders," npj Schizophrenia **7**(1), 25, doi: 10.1038/s41537-021-00154-3.

Tao, F., Esposito, A., et Vinciarelli, A. (**2020**). "Spotting the Traces of Depression in Read Speech : An Approach Based on Computational Paralinguistics and Social Signal Processing," dans *Interspeech 2020*, pp. 1828–1832, doi: 10.21437/Interspeech.2020-2888.

Tovar, A., Garí Soler, A., Ruiz-Idiago, J., Mareca Viladrich, C., Pomarol-Clotet, E., Rosselló, J., et Hinzen, W. (**2020**). "Language disintegration in spontaneous speech in Huntington's disease : a more fine-grained analysis," Journal of Communication Disorders **83**, 105970, doi: 10.1016/j.jcomdis.2019.105970.

Uher, R., Cumby, J., MacKenzie, L. E., Morash-Conway, J., Glover, J. M., Aylott, A., Propper, L., Abidi, S., Bagnell, A., Pavlova, B., Hajek, T., Lovas, D., Pajer, K., Gardner, W., Levy, A., et Alda, M. (**2014**). "A familial risk enriched cohort as a platform for testing early interventions to prevent severe mental illness," BMC Psychiatry **14**(1), 344, doi: 10.1186/s12888-014-0344-2.

Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., et Pantic, M. (**2013**). "AVEC 2013 - The Continous Audio/Visual Emotion and Depression Recognition Challenge," AVEC2013 .

van Bemmel, L., Harmsen, W., Cucchiarini, C., et Strik, H. (**2021**). "Automatic Selection of the Most Characterizing Features for Detecting COPD in Speech," dans *Speech and Computer*, édité par A. Karpov et R. Potapova, **12997** (Springer International Publishing), pp. 737–748.

Vásquez-Correa, J. C., Arias-Vergara, T., Rios-Urrego, C. D., Schuster, M., Rusz, J., Orozco-Arroyave, J. R., et Nöth, E. (**2019**). "Convolutional Neural Networks and a Transfer Learning Strategy to Classify Parkinson's Disease from Speech in Three Different Languages," dans *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*, édité par I. Nyström, Y. Hernández Heredia, et V. Milián Núñez, **11896** (Springer International Publishing, Cham), pp. 697–706, doi: 10.1007/978-3-030-33904-3_66.

Vázquez-Romero, A., et Gallardo-Antolín, A. (**2020**). "Automatic Detection of Depression in Speech Using Ensemble Convolutional Neural Networks," Entropy **22**(6), 688, doi: 10.3390/e22060688.

Verde, L., De Pietro, G., Ghoneim, A., Alrashoud, M., Al-Mutib, K. N., et Sannino, G. (**2021**a). "Exploring the Use of Artificial Intelligence Techniques to Detect the Presence of Coronavirus Covid-19 Through Speech and Voice Analysis," IEEE Access **9**, 65750–65757, doi: 10.1109/ACCESS.2021.3075571.

Verde, L., De Pietro, G., et Sannino, G. (**2021**b). "Artificial Intelligence Techniques for the Non-invasive Detection of COVID-19 Through the Analysis of Voice Signals," Arabian Journal for Science and Engineering doi: 10.1007/s13369-021-06041-4.

Vitale, F., Carbonaro, B., Cordasco, G., Esposito, A., Marrone, S., Raimo, G., et Verde, L. (**2021**). "A Privacy-Oriented Approach for Depression Signs Detection Based on Speech Analysis," Electronics **10**(23), 2986, doi: 10.3390/electronics10232986.

Vogel, A. P., Pearson-Dennett, V., Magee, M., Wilcox, R. A., Esterman, A., Thewlis, D., White, J. M., et Todd, G. (**2021**). "Adults with a history of recreational cannabis use have altered speech production," Drug and Alcohol Dependence **227**, 108963, doi: 10.1016/j.drugalcdep.2021.108963.

Voice and Speech Lab. "MEEI : Disordered Voice Database" .

Wang, B., Wu, Y., Taylor, N., Lyons, T., Liakata, M., Nevado-Holgado, A. J., et Saunders, K. E. (**2020**). "Learning to Detect Bipolar Disorder and Borderline Personality Disorder with Language and Speech in Non-Clinical Interviews," dans *Interspeech 2020*, pp. 437–441, doi: 10.21437/Interspeech.2020-3040.

Wang, B., Wu, Y., Vaci, N., Liakata, M., Lyons, T., et Saunders, K. E. A. (**2021**a). "Modelling Paralinguistic Properties in Conversational Speech to Detect Bipolar Disorder and Borderline Personality Disorder," dans *ICASSP 2021*, Toronto, ON, Canada, pp. 7243–7247, doi: 10.1109/ICASSP39728.2021.9413891.

Wang, H., Liu, Y., Zhen, X., et Tu, X. (**2021**b). "Depression Speech Recognition With a Three-Dimensional Convolutional Network," Frontiers in Human Neuroscience **15**, 713823, doi: 10.3389/fnhum.2021.713823.

Wang, J., Lv, K., Liu, C., Nie, X., Gowda, D., et Luan, S. (**2021**c). "Automatic Assessment for Severe Self-Reported Depressive Symptoms Using Speech Cues," IEEE Transactions on Cognitive and Developmental Systems **13**(4), 875–884, doi: 10.1109/TCDS.2020.3002512.

Wei, M., Du, J., Wang, X., Lu, H., Wang, W., et Lin, P. (**2021**). "Voice disorders in severe obstructive sleep apnea patients and comparison of two acoustic analysis software programs : MDVP and Praat," Sleep & Breathing = Schlaf & Atmung **25**(1), 433–439, doi: 10.1007/s11325-020-02102-4.

Woisard, V., Astésano, C., Balaguer, M., Farinas, J., Fredouille, C., Gaillard, P., Ghio, A., Giusti, L., Laaridh, I., Lalain, M., Lepage, B., Mauclair, J., Nocaudie, O., Pinquier, J., Pouchoulin, G., Puech, M., Robert, D., et Roger, V. (**2021**). "C2SI corpus : a database of speech disorder productions to assess intelligibility and quality of life in head and neck cancers," Language Resources and Evaluation **55**(1), 173–190, doi: 10.1007/s10579-020-09496-3.

Woldert-Jokisz, B. (**2007**). "Saarbruecken Voice Database" .

Xia, T., Spathis, D., Brown, C., Ch, J., Grammenos, A., Han, J., Hasthanasombat, A., Bondareva, E., Dang, T., Floto, A., Cicuta, P., et Mascolo, C. (**2021**). "COVID-19 Sounds : A Large-Scale Audio Dataset for Digital Respiratory Screening," dans *[Unpublished]*.

Yadav, S., Keerthana, M., Gope, D., Maheswari K., U., et Kumar Ghosh, P. (**2020**). "Analysis of Acoustic Features for Speech Sound Based Classification of Asthmatic and Healthy Subjects," dans *ICASSP 2020*, pp. 6789–6793, doi: 10.1109/ICASSP40776.2020.9054062.

Yamada, Y., Shinkawa, K., Kobayashi, M., Caggiano, V., Nemoto, M., Nemoto, K., Arai, T., et König, A. (**2021**a). "Combining Multimodal Behavioral Data of Gait, Speech, and Drawing for Classification of Alzheimer's Disease and Mild Cognitive Impairment," Journal of Alzheimer's Disease **84**(1), 315–327, doi: 10.3233/JAD-210684.

Yamada, Y., Shinkawa, K., Kobayashi, M., Nishimura, M., Nemoto, M., Tsukada, E., Ota, M., Nemoto, K., et Arai, T. (**2021**b). "Tablet-Based Automatic Assessment for Early Detection of Alzheimer's Disease Using Speech Responses to Daily Life Questions," Frontiers in Digital Health **3**, 653904, doi: 10.3389/fdgth.2021.653904.

Yamamoto, M., Takamiya, A., Sawada, K., Yoshimura, M., Kitazawa, M., Liang, K.-C., Fujita, T., Mimura, M., et Kishimoto, T. (**2020**). "Using speech recognition technology to investigate the association between timing-related speech features and depression severity," PloS One **15**(9), e0238726, doi: 10.1371/journal.pone.0238726.

Yan, Y., Mao, Y., Shen, Z., Wei, Y., Pan, G., et Zhu, J. (**2021**). "A High-Efficiency Fatigued Speech Feature Selection Method for Air Traffic Controllers Based on Improved Compressed Sensing," Journal of Healthcare Engineering **2021**, 1–10, doi: 10.1155/2021/2292710.

Yang, L., Jiang, D., et Sahli, H. (**2020**). "Feature Augmenting Networks for Improving Depression Severity Estimation From Speech Signals," IEEE Access **8**, 24033–24045, doi: 10.1109/ACCESS.2020.2970496.

Zhang, L., Duvvuri, R., Chandra, K. K. L., Nguyen, T., et Ghomi, R. H. (**2020**a). "Automated voice biomarkers for depression symptoms using an online cross-sectional data collection initiative," Depression and Anxiety **37**(7), 657–669, doi: 10.1002/da.23020.

Zhang, X., Ma, J., Li, Y., Wang, P., et Liu, Y. (**2021**). "Few-shot learning of Parkinson's disease speech data with optimal convolution sparse kernel transfer learning," Biomedical Signal Processing and Control **69**, doi: 10.1016/j.bspc.2021.102850.

Zhang, Y., Hu, W., et Wu, Q. (**2020**b). "Autoencoder Based on Cepstrum Separation to Detect Depression from Speech," dans *Proceedings of the 3rd International Conference on Information Technologies and Electrical Engineering*, ACM, Changde City Hunan China, pp. 508–510, doi: 10.1145/3452940.3453038.

Zhao, Y., Liang, Z., Du, J., Zhang, L., Liu, C., et Zhao, L. (**2021**). "Multi-Head Attention-Based Long Short-Term Memory for Depression Detection From Speech," Frontiers in Neurorobotics **15**, 684037, doi: 10.3389/fnbot.2021.684037.

Zhao, Z., Bao, Z., Zhang, Z., Cummins, N., Wang, H., et Schuller, B. (**2020**a). "Hierarchical Attention Transfer Networks for Depression Assessment from Speech," dans *ICASSP 2020*, Barcelona, Spain, pp. 7159–7163, doi: 10.1109/ICASSP40776.2020.9053207.

Zhao, Z., Bao, Z., Zhang, Z., Deng, J., Cummins, N., Wang, H., Tao, J., et Schuller, B. (**2020**b). "Automatic Assessment of Depression From Speech via a Hierarchical Attention Transfer Network and Attention Autoencoders," IEEE Journal of Selected Topics in Signal Processing **14**(2), 423–434, doi: 10.1109/JSTSP.2019.2955012.

Zhao, Z., Li, Q., Cummins, N., Liu, B., Wang, H., Tao, J., et Schuller, B. W. (**2020**c). "Hybrid Network Feature Extraction for Depression Assessment from Speech," dans *Interspeech 2020*, pp. 4956–4960, doi: 10.21437/Interspeech.2020-2396.

Deuxième partie

La somnolence : définitions et mesures

Résumé

Alors que le sens du terme « somnolence » est rarement défini dans les études l'estimant à partir d'enregistrements vocaux, les médecins du sommeil, à travers les discussions que nous avons pu avoir avec eux, font apparaître une richesse et une grande diversité de phénomènes associés à ce mot. Cette partie propose d'étudier la *somnolence*, à travers trois méthodes.

Tout d'abord, dans le chapitre 3, nous proposons deux approches naïves de la définition de ce terme. Dans un premier temps, nous utilisons les définitions proposées par trois dictionnaires de référence de la langue française. Dans un deuxième temps, nous proposons dans ce même chapitre le résultat obtenu par deux techniques de fouille de texte sur un grand corpus d'articles scientifiques en langue anglaise.

Dans une troisième approche, nous proposons d'étudier le concept de *somnolence* par le biais des outils de mesure qui ont été conçus pour le mesurer. Une revue générale (c'est-à-dire une revue de revues) des outils de mesure de la somnolence est ainsi proposée dans le chapitre 4.

Enfin, nous proposons dans le chapitre 5 les définitions et un schéma relationnel des différents construits pouvant être reliés au terme de *somnolence*, en nous basant sur la littérature précédemment publiée et les résultats obtenus dans les chapitres précédents.

Mots-clés

Somnolence ; Somnolence excessive ; Outils de mesure de la somnolence ; Revue générale ; Fouille de texte

Publication associée

Martin, V. P., Lopez, R., Dauvilliers Y., Rouas, J.-L., Philip, P., et Micoulaud-Franchi J.-A. (2022). Sleepiness in adults : An umbrella review of a complex construct. [En cours d'évaluation par les pairs].

Chapitre 3

De quoi la *somnolence* est-elle le nom ?

Sommaire

3.1 Contexte et motivations

Le langage que nous utilisons dans la vie courante – et plus particulièrement dans ce chapitre le vocabulaire – structure notre pensée (Bloom et Keil, 2001) et nous donne un « système d'opération pour exercer d'autres activités. » (Giere, 2010, p. 84) [1]

Une première idée pour aborder la définition de la somnolence est donc d'aller chercher dans des dictionnaires les définitions courantes de la *somnolence*. Cette approche est présentée dans la section 3.2.

Une deuxième approche proposée dans la section 3.3 de ce chapitre repose sur la fouille d'un grand corpus de textes et la représentation, sous forme de nuage de mots, des associations de concepts trouvés dans ce corpus.

Enfin, une autre approche basée sur la fouille de texte – mais cette fois-ci en utilisant une analyse en réseau – est présentée dans la section 3.4.

3.2 Approche naïve : définitions en langue courante

L'approche la plus naïve pour définir ce qu'est la *somnolence* est de chercher la définition de ce terme dans des dictionnaires de la langue française. Dans cette section, nous nous référons aux versions en ligne des dictionnaires Larousse, Robert et au dictionnaire de l'Académie Française, dans leur version du 4 mars 2022.

Par ailleurs, les mots *fatigue* et *somnolence* sont utilisés de manière *a priori* interchangeable dans le vocabulaire courant – par exemple dans l'expression « je suis fatigué ». Nous recherchons donc également la définition du terme *fatigue* dans ces dictionnaires, afin de compléter les précédentes définitions.

3.2.1 Définition de la *somnolence*

Dans les trois dictionnaires, la somnolence est définie comme un « état intermédiaire entre la veille et le sommeil. » Il convient donc de réitérer la recherche pour les deux termes de *veille* et *sommeil*.

Définition de la *veille*

La veille est définie de manière légèrement différente dans les trois dictionnaires. Alors que dans le Robert, elle est définie comme « [l']État d'une personne qui ne dort pas », le Larousse lui propose une définition non basée sur l'état, mais sur le processus d'éveil : « Sortir du sommeil, cesser de dormir ». Le dictionnaire de l'Académie française rajoute aux précédentes définitions un concept d'*alerte*, définissant la veille comme « État du corps de l'homme ou de l'animal, dans lequel les sens sont en action, par opposition à *Sommeil*. ». Cet état d'alerte est repris par le Larousse dans une deuxième définition : « avoir l'intelligence vive et alerte ».

Le dictionnaire de l'Académie française et Le Robert proposent une deuxième définition, liée non pas au réveil, mais à « [l']action de veiller volontairement dans le temps habituellement consacré au sommeil. », définissant la veille comme un « moment sans sommeil pendant le temps normalement destiné à dormir. »

La veille aurait donc deux dimensions selon les définitions précédentes : d'une part, un état de non-sommeil ou de sortie de sommeil, durant lequel le niveau d'alerte est élevé ; et

1. *"In learning an everyday language, a human acquires a kind of operating system for engaging in other activities."*

d'autre part une privation volontaire de sommeil, sur le temps « normalement » consacré au sommeil.

Définition du *sommeil*

Dans la définition de la somnolence, la veille est opposée au sommeil. Dans le Robert, ce dernier est défini comme « [l']état d'une personne qui dort, caractérisé essentiellement par la suspension de la vigilance et le ralentissement de certaines fonctions. ». Le sommeil est donc caractérisé à la fois par un changement physiologique et une baisse de vigilance ou de réactivité. Cette définition correspond également à la définition donnée par le Larousse : « État physiologique périodique de l'organisme (notamment du système nerveux) pendant lequel la vigilance est suspendue et la réactivité aux stimulations amoindrie. » Cette dernière définition rajoute également une dimension qui n'est pas retrouvée dans les autres définitions : la notion de *périodicité*.

Enfin, le dictionnaire de l'Académie française ajoute une dimension de *santé* à la notion de sommeil : « Interruption momentanée de certaines fonctions de l'activité vitale qui se produit surtout la nuit et procure le repos. »

Il est également intéressant de noter que la définition de la somnolence introduit une opposition entre *sommeil* et *éveil*, alors que la définition du sommeil dans le Larousse propose comme contraire à celui-ci les termes de *veille*, mais aussi d'*éveil* et de *réveil*.

Compléments de définitions du mot *somnolence*

La somnolence serait donc un état entre, d'une part, un état de veille, durant lequel les niveaux d'alerte et de vigilance sont élevés ; et d'autre part le sommeil, caractérisé par une périodicité, un changement physiologique de l'organisme et une baisse de vigilance, de réactivité aux stimuli.

Les autres définitions du mot *somnolence* proposées par les trois dictionnaires nous apportent là aussi des subtilités à la première définition.

Le dictionnaire de l'Académie française précise la temporalité associée au terme : « État de somnolence ». Il s'agit donc pour ce dictionnaire d'un *état*, qui a donc une durée limitée et cadrée dans le temps.

Le Larousse apporte une dimension comportementale à la somnolence, en complétant la définition par : « état d'engourdissement, de passivité, d'inertie. » La somnolence ne serait donc pas qu'une question de vigilance et d'alerte, mais aussi physique, corporelle. Cela est repris dans la définition du Robert, qui rajoute « Au figuré : Inaction, mollesse. », appuyant l'aspect comportemental et statique de la somnolence.

Enfin, le Robert nous livre un complément de définition qui est en opposition avec l'idée que la somnolence soit un état, mais plutôt un processus : « Tendance irrésistible à s'assoupir. » La somnolence peut ainsi être définie à la fois comme un état intermédiaire entre deux autres états (la veille et le sommeil), mais également comme la tendance à aller vers le sommeil, qui préfigure ce que les cliniciens appellent *la propension à l'endormissement*.

3.2.2 Fatigue

Dans le langage courant, les termes de *somnolence* et *fatigue* sont utilisés de manière interchangeable (cf. chapitre 5 pour plus de détails). Cependant, il semblerait que ça ne soit pas le cas dans le contexte clinique. Nous avons donc cherché, dans les trois dictionnaires, une

définition de la *fatigue* afin d'en chercher des caractéristiques permettant de la différencier de la somnolence.

Les trois définitions s'entendent à définir la fatigue comme un état « provoqué par une trop grande dépense de forces due à un travail excessif » (Académie française), « consécutif à un effort prolongé à un travail physique ou intellectuel intense » (Larousse) et « dû à un effort excessif » (Robert).

Les définitions s'opposent cependant sur la nature même de ce que l'on appelle *fatigue*. Alors que le Larousse la définit comme un « État physiologique » et le Robert comme un « Affaiblissement physique », le dictionnaire de l'Académie française définit la fatigue comme un « État de lassitude, » et donc un état psychique. De plus, la définition du Robert donne une deuxième version de la définition qui semble contredire sa première version : « sensation pénible qui l'accompagne [l'affaiblissement physique] ». De même, le dictionnaire de l'Académie française est ambigu puisque les exemples donnés relèvent à la fois du physique et du psychique : « Fatigue musculaire, intellectuelle, nerveuse. »

Il semble donc que pour la fatigue il faille faire une différence entre l'état physiologique et le ressenti conséquent de cet état, différence qui n'était pas apparue dans la définition de la somnolence.

Enfin, la définition du Larousse ajoute un élément qui fait tendre la définition vers une approche clinique de la fatigue en précisant ses conséquences : la fatigue « se tradui[t] par une difficulté à continuer cet effort ou ce travail. »

Conclusion

D'après les définitions précédentes, la somnolence peut être définie à la fois comme un « état intermédiaire entre la veille et le sommeil », eux-mêmes caractérisés par des niveaux de vigilance et d'éveil différents. Cependant, l'autre définition qui a émergé comme étant la « tendance irrésistible à l'endormissement » semble plus se rapprocher de la définition de la somnolence telle qu'elle est utilisée par les cliniciens. Les définitions ont également permis de soulever une question de santé de sommeil et de normes sociales, puisque le sommeil « procure le repos » et une définition de la veille est le maintien de l'éveil « durant le temps normalement destiné à dormir » : c'est donc qu'il y a une norme sur les moments durant lesquels nous sommes censés dormir.

La fatigue, elle, apparaît comme la conséquence d'un effort excessif, mais il n'est pas clair si ce terme traduit un état physiologique ou psychique. Dans les deux cas, la fatigue se traduit par une difficulté voire une impossibilité à continuer l'effort qui l'a fait naître. Elle se différencie bien de la somnolence : les deux mots sont interchangés à tort dans leur utilisation, mais leurs définitions sont bien distinctes. Dans la suite de ce document, nous nous concentrerons donc sur la *somnolence*.

Les définitions proposées de la *somnolence* par les dictionnaires de langue française commune ne prennent pas en compte toute la richesse de ce concept : si l'on retrouve les oppositions entre état sur le continuum veille-sommeil et propension à l'endormissement, les questions à la fois de l'opposition entre état physiologique et évaluation subjective et de plaintes cliniques n'émergent pas de ces définitions. De plus, des termes comme *hypersomnolence* n'existent pas dans ces dictionnaires alors qu'ils font partie du langage courant du clinicien.

Nous proposons dans les deux prochaines sections deux approches basées sur la fouille de textes d'articles médicaux, en langue anglaise (langue dominante pour la publication d'articles sur ce sujet) : nous espérons ainsi faire émerger les associations de concepts liés au

terme « somnolence » dans le vocabulaire employé dans une grande base de données textuelle clinique.

3.3 Première approche de fouille de textes : nuage de mots

3.3.1 Méthode

Afin de faire émerger les grands concepts également évoqués dans les articles travaillant sur la somnolence, nous avons collecté à l'aide de la librairie Python `Bio` les titres et résumés des 10000 premiers articles de PubMed comprenant le terme "sleepiness"[2], indépendamment de l'année, de la revue de publication ou du type d'article.

Ensuite, grâce à la librairie `nltk`, nous avons effectué les opérations suivantes :

— enlevé la ponctuation ;

— enlevé les *stop words*, qui sont des mots très courants en langue anglaise et non spécifiques au domaine que nous visons ;

— pris en compte les différentes formes d'une même racine de mots grâce au système `WordNetLemmatizer`. Par exemple, le terme *corpora* sera ramené à sa racine *corpus* grâce à ce module.

Ensuite, nous traçons un nuage de mots grâce au module `wordcloud`, dans lequel chaque mot a une taille proportionnelle à sa représentation parmi tous les mots isolés et traités par `nltk`. Nous nous limitons aux 100 mots de plus de 2 lettres les plus représentés dans le corpus. Le nuage de mots résultant est représenté dans la figure 3.1, et les mots ainsi que leur nombre d'observations sont disponibles en Annexe A.

FIGURE 3.1 – Nuage de mots représentant les termes les plus représentés dans le titre et le résumé des 10000 premiers articles contenant le mot "sleepiness" dans la base de données bibliographique PubMed.

3.3.2 Résultats

Cette figure contient des termes de différentes natures reliés à la somnolence. Le terme prédominant ("sleep", c'est-à-dire « sommeil ») illustre bien le lien entre somnolence et som-

2. Nous choisissons dans ce document la traduction de *sleepiness* pour *somnolence*.

meil que nous avions déjà observé dans le chapitre précédent. Les autres termes peuvent être catégorisés suivant les champs lexicaux suivants :

— pathologie ($n = 11$, $N_{obs} = 68263$) : patient, symptom, treatment, disorder, risk, clinical, disease, severity, disturbance, syndrome, therapy ;
— mesure ($n = 10$, $N_{obs} = 43192$) : scale, index, method, questionnaire, Epworth, outcome, polysomnography, objective, subjective, assessed ;
— apnée du sommeil ($n = 9$, $N_{obs} = 33816$) : osa, apnea, obstructive, cpap, ahi, osas, pressure, positive, airway
— construits et pathologies adjacents ($n = 6$, $N_{obs} = 17794$) : fatigue, depression, performance, association, cognitive, anxiety ;
— population ($n = 3$, $N_{obs} = 7490$) : child, woman, population ;
— santé ($n = 3$, $N_{obs} = 14213$) : quality, duration, health ;
— maladies du sommeil ($n = 2$, $N_{obs} = 6603$) : insomnia, narcolepsy ;
— autres ($n = 57$)

Il apparaît ainsi que dans le cadre de publications scientifiques médicales, la principale préoccupation concernant la somnolence est son aspect pathologique. La deuxième préoccupation dominante – qui va également nous intéresser dans le prochain chapitre 4 – est la façon de mesurer la somnolence, notamment par le biais de questionnaires, ou encore une opposition qui se dessine entre mesures objectives et subjectives. À part quasi égale avec la mesure de la somnolence, les termes liés à l'apnée du sommeil sont très représentés, et montrent l'importance de l'association entre la somnolence en tant que symptôme et l'apnée du sommeil. Nous avons également retrouvé la mention explicite de deux autres pathologies du sommeil (insomnie et narcolepsie), que nous avons préféré mettre dans une catégorie distincte.

Dans un deuxième temps, les construits adjacents à la somnolence comme la fatigue, la dépression ou les troubles cognitifs sont également très représentés, de même que les deux indicateurs de santé du sommeil : qualité et durée.

Enfin, sur un troisième plan, les populations pédiatriques et féminines semblent nécessiter des précautions particulières concernant leur prise en charge, la façon de mesurer la somnolence ou les conséquences de celle-ci.

3.3.3 Discussion et conclusion

Cette approche de fouille de textes nous a permis de nous rapprocher des différentes nuances de ce qu'est la somnolence, notamment grâce au changement de registre de langage (langage médical), et grâce à la puissance des outils computationnels qui permettent la fouille de très grandes bases de données de manière automatisée.

En revanche, deux limitations principales de cette étude freinent l'établissement d'une définition générale de la somnolence à partir des données précédentes. Tout d'abord, les mots isolés identifiés peuvent changer de sens suivant le contexte dans lequel ils sont employés, et si les grandes tendances se dessinent à travers des termes clés, une étude plus fine est nécessaire à l'établissement de la nature précise des liens entre ceux-ci. De plus, les catégories n'ont pas été induites à partir de ces mots isolés de leur contexte, mais à partir de connaissances préalables sur le domaine.

Dans la prochaine section, nous proposons un outil permettant d'éviter les deux précédents écueils : l'analyse en réseau.

3.4 Deuxième approche de fouille de textes : analyse en réseau

Nous proposons dans cette section une deuxième approche de fouille de textes, basée sur l'analyse en réseau. Cette approche revêt trois avantages par rapport à la précédente :

— Le graphe est construit sur des unités de plusieurs mots plutôt qu'un mot singulier[3] ;
— Les catégories sont induites à partir des données textuelles à analyser, et non construites a posteriori comme c'était le cas dans la partie précédente.
— Les relations entre les termes sont représentées par des liens proportionnels à la force de la liaison entre eux, permettant de déterminer une structure sous-jacente et d'expliciter les liens entre les différents groupes de termes.

L'approche utilisée a été développée et implémentée par Christophe Gauld (CCA – Hospices Civils de Lyon, doctorant en philosophie – Université Paris 1 Panthéon-Sorbonne) dans (Gauld et Micoulaud-Franchi, 2021a).

3.4.1 Méthode

La méthode d'obtention de la figure 3.2 est détaillée dans le précédent article. Nous en donnons ici les grandes lignes.

Après avoir collecté les titres, résumés et mots-clés (MeSH) de tous articles en anglais de la base de données PubMed qui contiennent le terme "sleepiness" sans restriction de date ou de type de publication, un post-traitement permet d'exclure les principaux synonymes.

Un algorithme de segmentation de texte est ensuite appliqué, permettant d'identifier et d'extraire des séquences grammaticales. De même que pour la création du nuage de mots, les termes extraits sont regroupés dans une même classe lorsque leurs racines sont identiques (*stemmed-version*). De plus, une ontologie systématique (en utilisant la hiérarchie des termes MeSH de PubMed/MEDLINE) permet généralement d'atteindre un niveau sémantique plus approfondi que la simple étude de la distribution du lexique dans un document, et a été appliquée au corpus d'articles.

L'analyse proposée utilise une analyse sémantique latente (*Latent Semantic Analysis*, LSA) – et plus particulièrement une allocation de Dirichlet latente (*Latent Dirichlet Allocation*)[4] – permettant d'établir des relations entre les termes en produisant un ensemble de concepts latents. Autrement dit, les termes apparaissant ensemble avec les mêmes distributions dans une partie d'un texte seront supposés avoir un sens proche, et seront regroupés dans la même catégorie.

À partir de cette analyse, un réseau lexical est construit, permettant de visualiser les termes les plus proches, c'est-à-dire les plus fréquemment associés. Les principales dimensions sémantiques issues de la fouille de données sont visualisées sous la forme d'un réseau lexical non orienté.

3.4.2 Résultats et discussion

Le graphe ainsi obtenu est représenté dans la figure 3.2. Nous avons identifié 7 sous-réseaux appartenant à 6 sous-domaines, dont nous proposons une analyse.

3. L'approche en nuage de mots aurait également pu autoriser les groupes de mots, mais elle n'offre que peu d'avantages par rapport à la version à un seul mot, et nous avons préféré limiter notre première approche sur les seuls termes à un seul mot

4. l'avantage de cette technique est qu'elle ne suppose pas une distribution gaussienne des termes, mais une distribution des densités

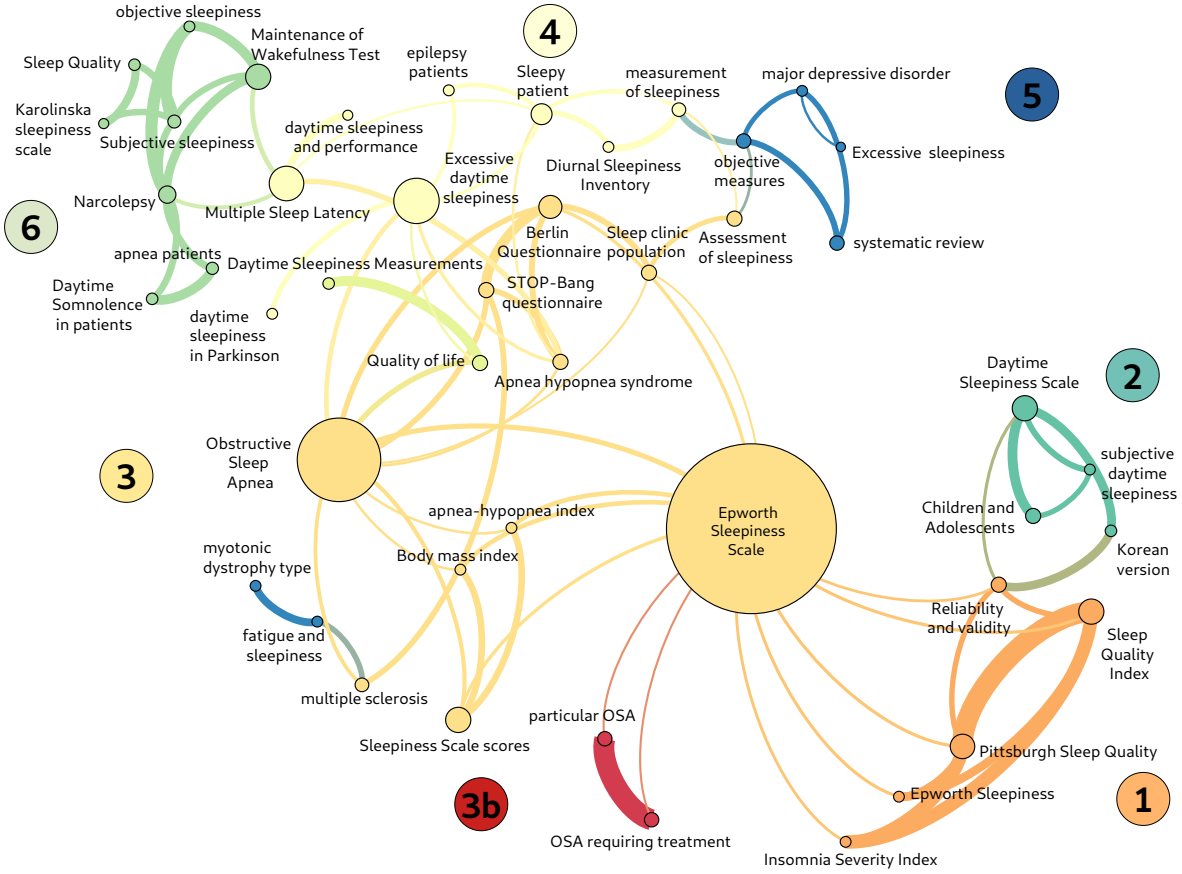

FIGURE 3.2 – Réseau lexical représentant les relations entre les différents termes prédominants dans les titres, résumés et mots-clés de tous les articles contenant le terme “sleepiness” dans la base de données bibliographique PubMed.

Epworth Sleepiness Scale

Avant d'explorer les six domaines identifiés, nous souhaitons tout d'abord souligner sur cette figure la place prépondérante que prend l'échelle de somnolence d'Epworth [ESS, (Johns, 1991)], outil favori des cliniciens travaillant dans les services de médecine du sommeil.

Exploration des 6 sous-domaines

(1) Échelles cliniques Son utilisation massive dans les services de médecine du sommeil explique notamment sa connectivité avec la zone orange, qui contient les trois questionnaires les plus utilisés lors des entretiens cliniques en médecine du sommeil : l'échelle de somnolence d'Epworth, l'index de sévérité de l'insomnie [ISI, (Bastien *et coll.*, 2001)], et l'index de qualité du sommeil de Pittsburgh [PSQI, (Buysse *et coll.*, 1989)]. Ces trois questionnaires sont souvent proposés ensemble, ce qui explique leur appartenance au même sous-réseau.

(2) Population particulière et influence de la langue Un petit sous-réseau turquoise contient des dimensions peu présentes dans le reste du graphe, mais malgré tout importantes pour ces questionnaires. D'une part, ce sous-réseau fait référence à la population à laquelle ils sont proposés, puisque ces questionnaires sont validés sur des populations adultes et nécessitent des adaptations ou des échelles à part pour les populations pédiatriques. D'autre part, ce sous-réseau contient également les problématiques liées à la langue de passation : si les concepts peuvent se traduire relativement facilement dans certaines langues (par exemple du français vers l'espagnol ou l'inverse), d'autres traductions sont plus ardues et nécessitent un travail plus profond (par exemple la traduction de l'anglais vers le coréen, illustré sur la figure par le noeud « version coréenne »).

(3) SAOS et forme particulière de SAOS Si l'on revient au noeud représentant l'ESS, son utilisation semble particulièrement liée au Syndrome Obstructif d'Apnée du Sommeil (SAOS, *OSA* en anglais), comme le montre sa forte connectivité avec les zones n°3b (SAOS particuliers) et n°3, dont le centre de sous-graphe est le noeud 'SAOS'.

(4) Somnolence Diurne Excessive, propension à l'endormissement, TILE Le sous-réseau correspondant au SAOS est lui-même connecté par de nombreux liens au sous-réseau n°4, dont les noeuds prédominants sont la 'somnolence diurne excessive', de la 'propension à l'endormissement' , avec notamment le 'Test Itératif de Latence d'Endormissement' (*Multiple Sleep Latency Test* en anglais).

(5) Comorbidités Ce sous-réseau assez généraliste est relié au petit sous-réseau n°5, qui comprend les noeuds 'somnolence excessive' et 'syndrome dépressif majeur', illustrant le lien entre somnolence et dépression (et anxiété) que nous avions déjà observé dans la section précédente.

(6) Somnolence chez des patients souffrant de maladies du sommeil Enfin, le dernier sous-réseau attaché au concept général de la somnolence diurne excessive comprend les noeuds 'Test de Maintien de l'Éveil' (*Maintenance of Wakefulness Test*), 'narcolepsie', 'Somnolence Diurne chez des patients', ou encore 'échelle de somnolence de Karolinska' (*Karolinska Sleepiness Scale* en anglais – KSS) et pourrait représenter un sous-groupe d'étude de la somnolence chez des patients souffrant de maladies du sommeil autres que le SAOS.

Populations particulières

Il est également intéressant de pointer certaines feuilles (c.-à-d. noeuds terminaux) spécifiant pour quelques-uns des sous-réseaux précédents des populations particulières. C'est le cas par exemple des patients atteints de sclérose en plaques (sous-réseau n°3, en bleu foncé), des patients épileptiques (sous-réseau n°4) ou encore de la somnolence excessive des patients atteints de la maladie de Parkinson (sous-réseau n°4).

3.4.3 Conclusion

En conclusion, cette approche permet de retrouver de nombreux éléments des deux précédentes analyses : mesure objective vs subjective, populations particulières, somnolence "normale" vs somnolence excessive. Si la structure proposée par cette analyse – reflétant les groupes d'idées présents dans les articles scientifiques contenus dans PubMed – n'a pas permis de faire émerger de relations entre les différents construits autour de la somnolence (fatigue, alerte, vigilance ...), elle a cependant fait émerger certains outils qui semblent très utilisés dans la recherche et la pratique clinique de médecine du sommeil.

Cette prépondérance des outils dans les termes associés au construit nous inspire une troisième approche de la définition de la *somnolence* : nous proposons dans le prochain chapitre de retrouver et d'organiser les différents concepts autour de la somnolence à partir des outils (questionnaires et tests médicaux) conçus pour les mesurer

Chapitre 4

Comment mesurer la somnolence ? – une revue générale

Cette revue a été conduite en collaboration étroite avec Jean-Arthur Micoulaud-Franchi (MCU-PH Université de Bordeaux, CHU de Bordeaux – JAMF) et Régis Lopez (PH, CHU de Montpellier – RL).

L'étude du construit par ses outils de mesure

Une des limites majeures des trois approches proposées dans le chapitre précédent est que celles-ci n'ont pas fait apparaître les multiples dimensions liées à la somnolence : malgré des travaux récents sur la clarification des construits liés à la somnolence, en particulier la fatigue (Shen *et coll.*, 2006) et la vigilance (van Schie *et coll.*, 2021), une confusion demeure autour de l'utilisation du mot *somnolence*.

Puisque l'étude des outils d'évaluation permet d'accéder aux construits sous-jacents (Radder, 2006), nous proposons d'investiguer et d'affiner la notion de somnolence en inventoriant les différents outils conçus pour son évaluation. Afin de refléter la richesse des différentes approches de la somnolence et de préserver les subtiles variations propres à chaque outil de mesure, nous avons procédé à une revue naïve de ces différents outils. Ainsi, nous ne les avons pas étiquetés avec un construit spécifique basé sur l'expertise de spécialistes, mais nous faisons plutôt l'hypothèse que les construits sous-jacents apparaîtront dans la structure des outils de mesure identifiés. Nous nous référons ici à la somnolence dans sa définition la plus large possible afin de capturer toutes les nuances des outils d'évaluation mesurant ce vaste concept.

Revue générale

Faire une revue systématique exhaustive des outils d'évaluation de la somnolence représenterait cependant une tâche irréalisable, nécessitant l'examen de toute la littérature jamais produite sur ce sujet. Pour rendre cette revue possible sans perdre en généralité, nous procédons à une revue générale (umbrella review), c'est-à-dire une revue des revues, suivant une méthodologie rigoureuse et standardisée adaptée de la méthode standard pour les revues systématiques (Page *et coll.*, 2021), qui rassemble toutes les preuves des revues existantes sur un sujet pour en donner une vue d'ensemble de haut niveau (Grant et Booth, 2009).

Cette revue générale a pour but d'identifier tous les outils d'évaluation de la somnolence mentionnés dans les revues publiées précédemment et portant sur la mesure de la somnolence chez les sujets adultes, et de les organiser en grandes catégories et selon une perspective temporelle. Ce faisant, nous cataloguons toutes les mesures de la somnolence qui sont suffisamment importantes pour être mentionnées dans un article de revue, donc reconnues par la communauté médicale et de recherche sur le sommeil au sens large.

4.1 Méthode

4.1.1 Stratégie de recherche et critères de sélection

Cette revue se conforme aux dernières recommandations PRISMA [*Preferred Reporting Items for Systematic Reviews and Meta-Analyses* (Page *et coll.*, 2021)].

Identification

Nous avons recherché dans les bases de données d'entrées bibliographiques PubMed, Web of Science, et Scopus les revues et revues systématiques liées à des mesures de la somnolence, publiées entre 1950 et le 30 mars 2021. Nous avons combiné plusieurs mots-clés, de la forme (mots-clés liés à la somnolence) ET (mots-clés liés à la mesure). Les requêtes complètes comprenaient 7 mots-clés pour la somnolence et 7 mots-clés pour la mesure [1]. Nous avons limité notre recherche aux titres des articles, et le type de ceux-ci était systématiquement filtré pour n'inclure que des revues et revues systématiques. De plus, une recherche manuelle de références a permis d'identifier deux références pertinentes supplémentaires (Lammers *et coll.*, 2020; Pertenais *et coll.*, 2019).

Vérification

Nous avons effectué une recherche manuelle des listes bibliographiques récupérées. Les articles ont été initialement sélectionnés par les investigateurs (JAMF, RL et moi-même) sur la base du titre, du résumé, et de la revue dans laquelle ils ont été publiés. Seules les références en langue anglaise, française ou espagnole ont été retenues. Par la suite, les textes intégraux des articles potentiellement éligibles ont été filtrés par les mêmes investigateurs selon les critères suivants. Nous n'avons sélectionné que les revues ou revues systématiques traitant des mesures de la somnolence, indépendamment de la population étudiée. Les éditoriaux, correspondances et résumés qui n'ont pas été filtrés par les bases de données bibliographiques ont été exclus manuellement. Nous avons également exclu les revues se focalisant sur une pathologie pour laquelle la somnolence est une conséquence, mais pas le sujet central de l'article. Enfin, les articles traitant de populations pédiatriques ou des adolescents ont été exclus. Le même processus d'évaluation a été appliqué à chaque article. Les désaccords ont été résolus par consensus.

4.1.2 Extraction des données et analyse

L'extraction des outils de mesure de la somnolence mentionnés dans les revues incluses a été effectuée consensuellement par JAMF, RL et moi-même.

Selon l'analyse proposée par Hempel (1970) du concept de *définition opérationnelle* pour un terme donné, nous avons considéré un outil d'évaluation de la somnolence comme étant défini par « des critères objectifs au moyen desquels tout chercheur scientifique peut décider, pour tout cas particulier, si le terme [somnolence] s'applique ou non » [2], c'est-à-dire une modalité technique qui fournit une mesure quantifiable liée à la somnolence selon l'auteur de la revue incluse.

1. (hypersomnolence OR sleepiness OR wakefulness OR drowsiness OR somnolence OR sleepy OR drowsy) AND (measur* OR questionn* OR scale* OR evaluat* OR defini* OR assess* OR diagnostic)
2. "objective criteria by means of which any scientific investigator can decide, for any particular case, whether the term [sleepiness] does or does not apply"

4.1.3 Traitement des données

Nombre d'outils de mesure par revue

Pour chaque revue incluse, nous avons tout d'abord extrait l'année de sa publication et le nombre d'outils distincts d'évaluation de la somnolence qu'il mentionne. Nous avons calculé la moyenne, l'écart-type, la médiane, et les nombres minimum et maximum d'outils mentionnés dans l'ensemble des revues. Enfin, nous avons calculé le coefficient de corrélation (r de Pearson) entre l'année de publication de chaque revue et le nombre d'outils qu'elle mentionne.

Classification des mesures extraites

Pour classer les différents outils d'évaluation de la somnolence précédemment extraits, nous avons défini des catégories basées sur la nature des outils utilisés pour opérationnaliser la mesure de la somnolence. Nous avons conçu les catégories dans le cadre d'une théorie ancrée (*grounded theory*) : les catégories sont construites de manière inductive, par un processus d'agrégation itératif (Smith et Smith, 1977). En conséquence, lorsque suffisamment de mesures partageant des caractéristiques communes sont rassemblées, une nouvelle catégorie est créée et les critères d'appartenance de chaque catégorie sont mis à jour. Deux spécialistes de la médecine du sommeil (JAM et RL) ont proposé une classification pour chaque outil. En cas de divergence majeure entre les catégories, le désaccord a été résolu par consensus.

Métriques des articles originels

Nous avons associé un article originel à chaque outil d'évaluation de la somnolence, défini comme l'article scientifique le plus ancien mentionnant son utilisation pour évaluer un construit lié à la somnolence. Nous avons extrait l'année de publication de ces articles dont nous avons calculé le minimum, le maximum et la distribution sur les décennies correspondantes des années 1960 à 2010. Enfin, pour chaque catégorie, nous avons calculé l'année médiane de publication de l'article originel associé aux outils classés dans la catégorie.

Synthèse des données

Afin de faciliter l'appréhension de ces données, nous avons représenté chaque outil d'évaluation de la somnolence extrait sur un plan dirigé d'une part par l'année de publication de l'article originel associé ; et d'autre part par la catégorie dans laquelle l'outil d'évaluation a été précédemment classé. La taille de chaque point correspond au nombre de revues qui mentionnent les outils d'évaluation de la somnolence. Pour appréhender toutes les subtilités sous-jacentes aux données, tout en rendant la figure claire et lisible, nous avons affiché les données extraites sur un graphique à bulles interactif en utilisant la bibliothèque Python `Plotly` et l'API `Chartstudio`. L'utilisation d'un survol interactif permet d'afficher dynamiquement les informations précédentes et l'article de référence lors du survol des outils d'évaluation de la somnolence, et l'axe interactif permet d'effectuer un zoom avant et arrière à volonté.

4.2 Résultats

4.2.1 Sélection des revues

La figure 4.1 contient le diagramme PRISMA illustrant le processus d'inclusion et d'exclusion des articles de revue. Sur les 92 articles de revue identifiés après élimination des

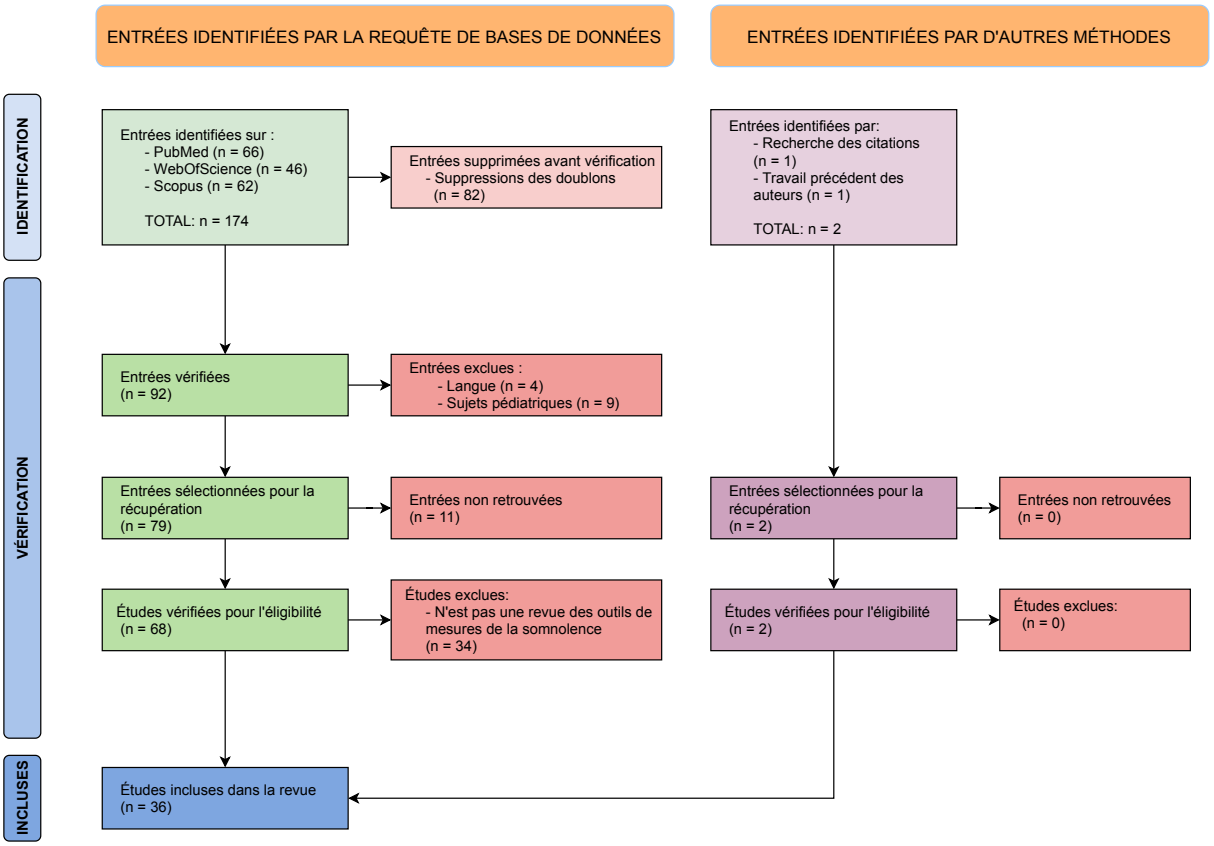

FIGURE 4.1 – Diagramme PRISMA illustrant le processus d'inclusion et d'exclusion des articles de revue filtrés.

doublons, 57 ont été exclus : 34 n'étaient pas des revues d'outils d'évaluation de la somnolence, 11 n'avaient pas de texte intégral disponible, 9 portaient sur une population pédiatrique et 4 n'étaient pas en anglais, français ou espagnol. 36 articles de revue ont donc été inclus. Une description complète des articles de synthèse inclus et exclus et des outils de mesure de la somnolence identifiés est disponible dans l'Annexe B .

4.2.2 Description des revues et des outils de mesure de la somnolence

Les 36 articles de revue ont été publiés entre 1982 et 2020, 2 dans les années 1980 (5.5%), 2 dans les années 1990 (5.5%), 10 dans les années 2000 (27.8%) et 22 dans les années 2010 (61.1%). Le nombre moyen de mesures de la somnolence mentionnées par revue était de 11.8 (écart-type : 8.0), avec une médiane de 10 (intervalle 1-39).

Nombre d'outils de mesure par catégorie

Quatre-vingt-dix-neuf outils d'évaluation de la somnolence ont été extraits des revues sélectionnées (cf. Annexe B). Ils ont été classés de manière consensuelle selon huit catégories :

1. Questionnaires ($n = 54, 54.5\%$) : instrument constitué d'une série de questions ou d'autres types de propositions visant à recueillir des informations auprès d'un répondant (le sujet lui-même ou un observateur).

2. Mesures dérivées de l'électroencéphalographie (EEG) ($n = 7, 7.1\%$) : mesures quantitatives basées sur les signaux EEG.
3. Mesures dérivées de la polysomnographie (PSG) ($n = 10, 10.1\%$) : mesures quantitatives basées sur une combinaison d'EEG, d'électrooculographie et d'électromyographie pour définir les états de sommeil ou de veille.
4. Mesures basées sur les performances ($n = 12, 12.1\%$) : mesures quantitatives liées aux performances cognitives ou psychomotrices sur une tâche spécifique.
5. Mesures de l'activité ($n = 7, 7.1\%$) : outils d'évaluation quantitative et qualitative basés sur les comportements et les mouvements observés.
6. Mesures liées aux yeux ($n = 3, 3.0\%$) : mesures quantitatives basées sur les mouvements des yeux et le clignement des paupières.
7. Mesures du système autonomique ($n = 4, 4.0\%$) : mesures quantitatives dérivées des signaux physiologiques liés à l'activation du système nerveux autonome.
8. Autres mesures ($n = 2, 2.0\%$) : mesures n'entrant pas dans l'une des catégories précédentes (c'est-à-dire mesures biologiques et mesures dérivées de la magnétoencéphalographie).

Nombre d'outils de mesure par revue

En moyenne, chaque outil de mesure de la somnolence a été mentionné 4.0 fois par revue (é-t : 6.1). Cinquante-neuf outils (59.6%) ont été mentionnés moins de trois fois, dont 38 (38.4%) une seule fois. Six outils ont été mentionnés par plus de la moitié des revues incluses : le test itératif de latence d'endormissement (TILE, $n = 31, 86.1\%$), le questionnaire de somnolence d'Epworth (ESS, $n = 29, 80.6\%$), le test de maintien de l'éveil (TME, $n = 26, 72.1\%$), l'échelle de somnolence de Stanford (SSS, $n = 25, 69.4\%$), l'échelle de somnolence de Karolinska (KSS, $n = 21, 58.3\%$) et le test psychomoteur de vigilance (PVT, $n = 18, 50.0\%$).

Aucune des revues incluses n'a mentionné plus de la moitié des outils de mesure de la somnolence que nous avons identifiés. Le nombre d'outils mentionnés dans une revue est positivement corrélé avec l'année de sa publication ($r = 0.37, p = 0.02$).

Trente-cinq revues (97.3%) mentionnent au moins un questionnaire, 32 d'entre elles (88.9%) mentionnent au moins une mesure dérivée de la PSG, 24 d'entre elles (66.7%) mentionnent au moins une mesure basée sur les performances, 17 d'entre elles (42.7%) mentionnent au moins une mesure du système autonomique, 16 d'entre elles (44.4%) mentionnent au moins une mesure dérivée de l'EEG, 15 d'entre elles (41.7%) mentionnent au moins une mesure de l'activité, et 12 d'entre elles (33.3%) mentionnent au moins une mesure de l'activité oculaire. Le nombre moyen de catégories par article est de 4.3 (é-t : 1.7)

Articles originels des outils de mesure de la somnolence, par année et par catégorie

Les articles originaux associés à chaque outil d'évaluation de la somnolence ont été publiés entre 1961 et 2019. Sept (7.1%) des articles originaux ont été publiés dans les années 1960, six (6.1%) dans les années 1970, 21 (21.2%) dans les années 1980, 37 (37.4%) dans les années 1990, 13 (13.1%) dans les années 2000 et 15 (15.1%) dans les années 2010. Le délai moyen entre l'année de publication de l'article original et l'année de sa première occurrence dans un article de synthèse était de 14.0 ans (é-t : 10.8).

Les articles originaux des questionnaires ont été publiés entre 1973 et 2019 (année médiane de publication : 1995), ceux des mesures basées sur l'EEG entre 1964 et 2008 (médiane : 1995),

ceux des mesures basées sur la PSG entre 1979 et 2018 (médiane : 1986), ceux des mesures autonomiques entre 1961 et 1982 (médiane : 1964), ceux des mesures du comportement et de l'activité motrice entre 1986 et 2001 (médiane : 1995), ceux des mesures oculaires entre 1984 et 1989 (médiane : 1987), et ceux des mesures basées sur la performance entre 1976 et 2010 (médiane : 1989).

Nous avons représenté dans la figure reffig :chartstudio les outils d'évaluation de la somnolence selon les huit catégories, l'année de publication de l'article originel et le nombre de mentions dans les articles de revue inclus. Une version interactive de cette figure est disponible en ligne.

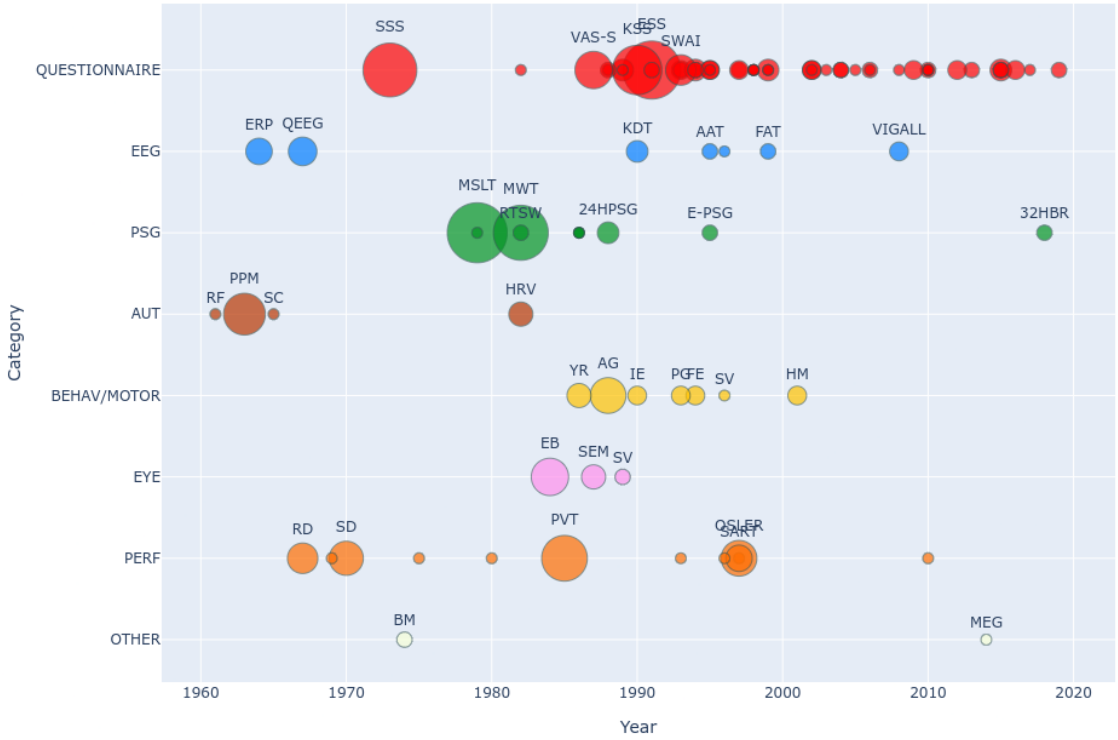

FIGURE 4.2 – 99 outils de mesure de la somnolence en fonction de l'année de publication du premier article mentionnant leur utilisation pour la somnolence et de la catégorie dans laquelle ils ont été classés. La taille des bulles est proportionnelle au nombre de fois que chaque outil a été cité dans les 36 revues incluses.

4.3 Discussion

Dans cette revue générale, nous avons identifié tous les outils de mesure mentionnés dans de précédentes revues portant sur les outils de mesure de la somnolence chez une population adulte, au cours des six dernières décades. Une centaine d'outils ont été identifiés, reflétant la remarquable ingéniosité de la communauté de recherche en médecine du sommeil pour développer des outils innovants mesurant ce symptôme central.

4.3.1 Principaux outils de mesure

Reflets de l'ensemble des outils de mesure

Les six outils d'évaluation les plus cités (ESS, TILE, TME, SSS, KSS et PVT) reflètent bien la grande diversité des 99 outils d'évaluation de la somnolence extraits de cette revue générale, et notamment les deux principaux contextes dans lesquels la somnolence est mesurée. Alors que l'ESS, le TILE et le TME sont traditionnellement utilisés dans un contexte clinique, la SSS, la KSS et le PVT évaluent principalement la somnolence dans des contextes expérimentaux.

En outre, ces six mesures reflètent également les différents concepts qui sous-tendent le phénomène de la somnolence : l'ESS et le TILE sont généralement considérés comme des mesures de la *propension à l'endormissement* (Carskadon et Dement, 1979; Arand *et coll.*, 2005) alors que, la SSS, la KSS et le PVT mesurent la *perception subjective de somnolence* (Åkerstedt *et coll.*, 2014; Dinges et Powell, 1985; Hoddes *et coll.*, 1973). Enfin, les conséquences fonctionnelles de la somnolence sont généralement évaluées avec l'ESS et le TME (Johns, 1991; Mitler *et coll.*, 1982), notamment en ce qui concerne les accidents de la route liés à la somnolence (Bioulac *et coll.*, 2017).

Absence de l'inertie du sommeil et de la quantité excessive de sommeil

Il est à noter que les autres dimensions du spectre de l'hypersomnolence (c'est-à-dire la quantité excessive de sommeil et l'inertie du sommeil, cf chapitre 5 suivant) ne sont pas explorées par ces six outils largement reconnus par la communauté clinique et de recherche sur le sommeil. Deux hypothèses peuvent sous-tendre cette observation.

Une première hypothèse est que les experts du sommeil différencient la quantité excessive de sommeil et l'inertie du sommeil de la somnolence excessive, ce qui signifie que les outils les mesurant sont décrits dans des articles qui n'entrent pas dans le cadre de notre revue générale, axée sur la somnolence excessive.

Deuxièmement, les outils mesurant la quantité excessive de sommeil et l'inertie du sommeil pourraient être sous-représentés parce qu'il s'agit de concepts individualisés plus récemment, et qu'il ne s'est pas écoulé assez de temps pour que les outils mesurant ces concepts soient largement adoptés dans les cliniques du sommeil et les paradigmes de recherche – et donc cités dans une revue. En ce sens, une étude récente a démontré la nécessité d'une évaluation standardisée de la quantité excessive de sommeil (Evangelista *et coll.*, 2018). D'autres études récentes ont montré que le PVT peut fournir une mesure quantitative de l'inertie du sommeil chez les patients souffrant d'hypersomnie idiopathique et d'autres troubles du sommeil (Evangelista *et coll.*, 2021; Trotti *et coll.*, 2022).

Questionnaires vs mesures dites "objectives"

La nature des six outils d'évaluation de la somnolence les plus cités illustre la distinction entre les questionnaires (ESS, SSS et KSS), principalement conçus pour refléter l'expérience subjective de la somnolence, et les mesures dites "objectives" de la somnolence (TILE, TME et PVT), souvent considérées comme des évaluations de référence ; avec des préoccupations majeures concernant la faible corrélation entre ces mesures (Kendzerska *et coll.*, 2014).

Les questionnaires représentent plus de la moitié (54.5%) de tous les outils d'évaluation de la somnolence mentionnés dans les revues incluses. La diversité de ces questionnaires souligne la complexité de l'expérience subjective de la somnolence. Certains outils, comme la SSS et la KSS, ont été développés pour mesurer la somnolence à un moment précis. Ces

questionnaires simples semblent être sensibles à la privation de sommeil et au moment de la journée durant lequel ils sont remplis (Herscovitch et Broughton, 1981; Kaida *et coll.*, 2006), et sont largement utilisés dans des contextes expérimentaux. Par ailleurs, d'autres questionnaires ont été conçus pour évaluer les conséquences de la somnolence dans la vie quotidienne. L'ESS, l'inventaire de somnolence diurne basé sur l'observation et l'entretien (Onen *et coll.*, 2016, ODSI) ou l'inventaire de l'activité veille-sommeil (Rosenthal *et coll.*, 1993, SWAI) entrent dans cette catégorie, et sont plus susceptibles d'être utilisés dans des contextes cliniques.

L'analyse des construits des questionnaires révèle également la nécessité d'étudier les dimensions symptomatiques spécifiques de la somnolence. La plupart des échelles, comme l'ESS, portent principalement sur la propension à l'endormissement. Certains questionnaires comme le test d'alerte de l'hôpital de Toronto (Shahid *et coll.*, 2016, THAT) ou l'échelle de résistance au sommeil (Violani *et coll.*, 2003, RSS) ont été conçus pour évaluer la capacité à rester éveillé dans diverses conditions. Quelques questionnaires évaluent l'expérience de la somnolence (SSS, KSS), l'inertie du sommeil (Kanady et Harvey, 2015, SIQ) ou la quantité excessive de sommeil (ODSI). Enfin, certains questionnaires ont été conçus pour évaluer spécifiquement les conséquences fonctionnelles de la somnolence, comme l'ESS ou le questionnaire de conséquences fonctionnelles du sommeil (Weaver *et coll.*, 1997, FOSQ). Cependant, la majorité des questionnaires que nous avons identifiés dans notre revue générale ne portent que sur une seule dimension de la somnolence, avec très peu d'exemples de questionnaires multidimensionnels.

Pour répondre à cette problématique, l'échelle de sévérité de l'hypersomnie idiopathique (Dauvilliers *et coll.*, 2019, IHSS) et l'indice de sévérité de l'hypersomnie (Kaplan *et coll.*, 2019, HSI) ont été développés presque simultanément à la fin des années 2010, dans le but d'évaluer l'ensemble du spectre de l'hypersomnolence et de ses conséquences fonctionnelles. Ceux-ci ont démontré d'excellentes propriétés psychométriques (Fernandez-Mendoza *et coll.*, 2021; Rassu *et coll.*, 2021) et ont donc le potentiel d'être largement utilisés pour évaluer l'hypersomnolence tant dans la pratique clinique que dans les essais contrôlés randomisés (Dauvilliers *et coll.*, 2022).

4.3.2 Nouveaux outils de mesure de la somnolence et *dépendance au chemin*

Dès 1993, Pickering a identifié un phénomène *dépendance au chemin*[3] de la production de nouvelles connaissances (Pickering, 1993, p.185) :

> Mon analyse suggère une dépendance à la *situation* et au *chemin* de la production de connaissances. D'une part, ce qui est reconnu comme connaissance empirique ou théorique, indépendamment de l'époque, est une fonction de l'espace matériel/conceptuel/disciplinaire/social, etc. dans lequel elle s'inscrit, et non pas juste du monde [tel qu'il est].[4]

Les nouveaux outils d'évaluation de la somnolence, malgré leurs performances et leur validation par une partie de la communauté du sommeil, tombent eux aussi sous le joug de cette dépendance au chemin, qui empêche leur adoption à grande échelle. Pour plus d'informations sur le phénomène de dépendance au chemin, nous redirigeons le lecteur vers (Pickering, 1993; Chu et Evans, 2021; Liebowitz et Margolis, 2014).

3. *path dependency*

4. "My analysis of practice [...] points to a situatedness and path dependence of knowledge production. On the one hand, what counts as empirical or theoretical knowledge at any time is a function not just of how the world is but of the specific material-conceptual-disciplinary-social-etc. space in which knowledge production is situated."

Cet effet de dépendance au chemin semble avoir conduit à la sursélection de certains outils au détriment d'autres, comme l'illustrent nos résultats, qui montrent que 38 mesures (38,4%) ne sont mentionnées qu'une seule fois dans les articles de revue inclus. Différents mécanismes sous-jacents à cet effet peuvent être identifiés. Tout d'abord, comme dans de nombreuses autres disciplines, on peut observer un cycle où les premiers outils sont, au fil du temps, parmi les plus utilisés, et donc, les plus validés, étant eux-mêmes largement adoptés par la communauté. De plus, cette dépendance au chemin pourrait également être le résultat de choix historiques des sociétés scientifiques et académiques qui ont structuré la récente discipline de la médecine du sommeil. La surreprésentation de deux outils de mesure de la somnolence basés sur la PSG (TILE et TME) est probablement liée au rôle central de la polysomnographie dans la pratique de la médecine du sommeil (Gauld et Micoulaud-Franchi, 2021b). Selon l'*American Board of Family Medicine*, la capacité à utiliser la PSG même définirait ce qu'est un médecin du sommeil : « un expert [...] compétent dans l'analyse et l'interprétation de la polysomnographie, bien renseigné sur les champs de recherche émergents, et la gestion d'un laboratoire de sommeil »[5].

Le TILE est également un test essentiel en médecine du sommeil, ses résultats étant des critères de diagnostic depuis la première classification internationale des troubles du sommeil [ICSD-I, 1990] et la quatrième révision du Manuel diagnostique et statistique des troubles mentaux [DSM-IV, 1994]. L'adoption à grande échelle du TILE, mais aussi du TME, a également été largement influencée par l'établissement et les mises à jour ultérieures des directives des critères pratiques par l'*American Academy of Sleep Medicine*, contrairement à la plupart des autres outils d'évaluation de la somnolence [par exemple en 2005 (Littner *et coll.*, 2005) ou plus récemment en 2021 (Krahn *et coll.*, 2021)]. La plupart des consensus d'experts ayant formulé des recommandations sur l'utilisation d'outils d'évaluation de la somnolence concernaient des tests objectifs, et rarement des questionnaires. Cependant, nos résultats suggèrent que ce phénomène affecte également les questionnaires, comme par exemple dans le domaine des essais thérapeutiques, où les résultats doivent être évalués à l'aide d'outils d'évaluation subjectifs fiables et comparables entre les études.

4.3.3 Intégration dans l'histoire de la médecine du sommeil

Nos résultats suggèrent que le développement et l'adoption d'outils d'évaluation de la somnolence s'intègrent dans l'histoire de la médecine du sommeil. Nous proposons ici d'interpréter la figure 4.2 au regard des six dernières décennies de développement de recherche sur le sommeil.

Avant les années 1970, l'approche psychophysiologique semble avoir été le moteur du développement des premiers outils d'évaluation de la somnolence, avec le souci de trouver des marqueurs physiologiques de la somnolence en tant qu'état de faible niveau d'activation, tel que l'EEG (Daniel, 1967; Haider *et coll.*, 1964), la fréquence respiratoire (Crawford, 1961), la pupillométrie (Lowenstein *et coll.*, 1963) ou la conductance cutanée (Davies et Krkovic, 1965). À cette époque, les premières mesures liées à la performance ont été développées pour évaluer les conséquences fonctionnelles de la somnolence. Il s'agit notamment des premières évaluations de la somnolence avec des tests de performance de conduite en situation réelle ou simulée (Brown *et coll.*, 1967; Heimstra, 1970), qui ont interpellé très tôt la communauté, avant

5. "A family physician with demonstrated expertise in the diagnosis and management of clinical conditions that occur during sleep, that disturb sleep, or that are affected by disturbances in the wake-sleep cycle. This specialist is skilled in the analysis and interpretation of comprehensive polysomnography, and well versed in emerging research and management of a sleep laboratory."

même que le risque d'accidents liés à la somnolence ne devienne une préoccupation de santé publique.

Au cours des années 1970, peu d'outils d'évaluation de la somnolence ont été développés. Cependant, cette décennie a introduit un changement de paradigme et ouvert la voie au domaine de la psychométrie, qui allait devenir une approche essentielle de l'évaluation de la somnolence au cours des décennies suivantes, avec l'introduction, en 1973, de l'échelle de somnolence de Stanford (Hoddes *et coll.*, 1973, SSS).

Les deux tests objectifs les plus utilisés pour mesurer la somnolence (TILE et TME) ont été développés dans les années 1980 : avec l'introduction de la PSG comme mesure de la somnolence, cette décennie a vu l'émergence du concept de latence d'endormissement comme mesure continue de la propension à l'endormissement. Développé en 1985, le PVT (Dinges et Powell, 1985) a simplifié l'évaluation des conséquences de la somnolence sur la performance, via la mesure de processus cognitifs plus purs comme l'attention. Cette dernière est la mesure principale d'autres tests importants développés au cours de la décennie suivante, comme le test de résistance au sommeil d'Oxford (Bennett *et coll.*, 1997, OSLER) ou la tâche de réponse à l'attention soutenue (Robertson *et coll.*, 1997, SART).

La publication de l'index de qualité du sommeil de Pittsburgh en 1989 (Buysse *et coll.*, 1989, PSQI) – en tant que mesure des troubles du sommeil – et de l'ESS en 1991, ont ouvert la voie à l'ère de la psychométrie dans le domaine de l'évaluation de la somnolence. Depuis les années 1990, la somnolence n'est plus seulement perçue comme un état physiologique, mais aussi comme un symptôme clinique. Ainsi, au cours des années 2000, et surtout des années 2010, une attention particulière a été portée aux autres dimensions de la somnolence, telles que l'inertie du sommeil ou la quantité excessive de sommeil, qui sont les principaux symptômes des troubles centraux de l'hypersomnolence. Au cours de ces décennies, les innovations thérapeutiques dans le domaine des troubles du sommeil ont créé un fort besoin d'outils spécifiques à la maladie et sensibles aux changements à utiliser dans les essais contrôlés randomisés (Dauvilliers *et coll.*, 2022; Ingravallo *et coll.*, 2020).

4.3.4 Tendances récentes

D'autres tendances, notamment au cours des deux dernières décennies, ont pu échapper à notre analyse en raison des limites propres aux revues générales (Grant et Booth, 2009). En effet, le délai moyen entre la première apparition d'une mesure de la somnolence dans la littérature scientifique et sa première mention dans un article de synthèse est de 14 ans. Par conséquent, certains outils d'évaluation de la somnolence peuvent avoir été développés trop récemment pour apparaître dans cette revue générale.

À titre d'exemple, une nouvelle approche dans le développement des outils d'évaluation de la somnolence semble émerger depuis la fin des années 2010 et propose d'affiner les outils d'évaluation existants en collectant des mesures alternatives aux tests sans en changer la procédure, comme cela a été fait récemment pour la PSG (Lim *et coll.*, 2020). Par exemple, l'analyse du micro-sommeil (Anniss *et coll.*, 2021; Des Champs de Boishebert *et coll.*, 2021; Hertig-Godeschalk *et coll.*, 2020; Morrone *et coll.*, 2020) ou l'analyse vidéo des mouvements oculaires (Kratzel *et coll.*, 2021) peuvent notamment améliorer l'évaluation du risque de conduite lié à la somnolence, en plus des mesures conventionnelles de latence d'endormissement. Dans le même ordre d'idée, sur la base du TILE, d'autres études ont identifié des seuils alternatifs (Pizza *et coll.*, 2019) ou des mesures alternatives [comme par exemple, la dynamique de transition veille-sommeil (Drakatos *et coll.*, 2013; Kawai *et coll.*, 2020) ou le pourcentage de sommeil paradoxal (Murer *et coll.*, 2017)] comme des mesures pertinentes pour mieux carac-

tériser la somnolence excessive dans la narcolepsie.

En utilisant des tests standardisés basés sur la PSG, d'autres études se sont également intéressées à la conciliation entre la conscience qu'ont les sujets de leur somnolence et les paramètres standard de somnolence basés sur la PSG. C'est le cas de l'indice de somnolence de Barcelone (Guaita *et coll.*, 2015, BSI), qui est un questionnaire dont la méthodologie de conception même était basée sur la PSG, le TME, le TILE et le SART, générant ainsi un questionnaire avec une excellente validité externe sur ces mesures. Cependant, la validité du BSI en milieu clinique reste à prouver, cette étude manquant de reproductibilité.

Plus récemment, deux études se sont intéressées à l'autoperception des sujets passant un TME. La première a montré l'impact de la maladie de Parkinson sur la conscience du moment d'endormissement (Bargiotas *et coll.*, 2019) tandis que la seconde a montré que la mauvaise perception du début du sommeil pendant le test est liée au risque d'accidents de la route dus à la somnolence (Sagaspe *et coll.*, 2021).

Enfin, un travail récent a étudié la relation entre les temps de sommeil mesurés par la PSG et leur équivalent perçu par les patients (Valko *et coll.*, 2021), trouvant une influence significative du type de trouble du sommeil et de l'éveil sur la caractéristique de la mauvaise perception (sous- ou surestimation) du temps de sommeil total. Malgré ces efforts, les études portant sur les variables prédictives de la conscience que les sujets ont de leur somnolence sont peu nombreuses. Compte tenu du potentiel que cela représente, notamment dans le domaine critique de la sécurité routière (Cai *et coll.*, 2021), nous pensons que des efforts supplémentaires devraient être faits dans ce sens.

4.3.5 Limitations

Comme toutes les revues générales, les analyses et les conclusions proposées s'appuient sur les revues publiées précédemment qui sont référencées dans les bases de données qui ont été utilisées (Grant et Booth, 2009). En conséquence, ce travail n'a pas pu rapporter tous les outils d'évaluation de la somnolence existants et certaines catégories ne sont pas apparues dans cette revue. Il s'agit notamment des mesures de la somnolence basées sur des données de neuroimagerie fonctionnelle (Dauvilliers *et coll.*, 2017a; Gool *et coll.*, 2020; Vallat *et coll.*, 2019), sur la génétique (Honda *et coll.*, 2018; Tanida *et coll.*, 2021; Wang *et coll.*, 2019) ou sur des approches biologiques (Esfandyarpour *et coll.*, 2019; Miyagawa *et coll.*, 2011; Pajcin *et coll.*, 2017). De plus, les critères de sélection utilisés pour cette revue générale ont exclu les outils d'évaluation de la somnolence spécifiquement développés pour la population pédiatrique, ou pour des maladies spécifiques (Chaudhuri *et coll.*, 2002; Hermans *et coll.*, 2013). Néanmoins, l'utilisation d'une revue générale a permis de capturer tous les outils d'évaluation adoptés par la communauté du sommeil et est utile pour développer des recommandations pour la pratique clinique et la recherche (Grant et Booth, 2009).

Deuxièmement, en raison de l'absence d'une structure organisationnelle commune des outils dans les différentes revues incluses, nous n'avons pas été en mesure de catégoriser systématiquement ces mesures de la somnolence en fonction des dimensions symptomatiques ou des mécanismes psychophysiologiques qu'elles exploraient (c'est-à-dire la propension au sommeil, la somnolence, la vigilance, la quantité excessive de sommeil, l'inertie du sommeil, la capacité à maintenir l'éveil...). Bien que des articles récents aient proposé de clarifier certains construits connexes – tels que la fatigue (Shen *et coll.*, 2006), la quantité excessive de sommeil (Lammers *et coll.*, 2020), ou la vigilance, l'alerte et l'éveil (van Schie *et coll.*, 2021), ce que l'on appelle *somnolence* reste un construit flou. Néanmoins, ce travail est un premier pas vers une définition consensuelle des différentes dimensions de la somnolence en proposant

un cadre plus clair pour les discussions futures entre les différents acteurs de la médecine du sommeil.

Troisièmement, une autre limite est liée à la méthodologie d'extraction des données. En effet, les construits explorés n'ont pas toujours été précisés par les auteurs et les descriptions verbales employées dans les articles de revue étaient très diversifiées. Des approches systématiques et à plus grande échelle seraient probablement plus appropriées pour explorer et classer ces outils en fonction des concepts explorés. Parmi les analyses envisagées, les techniques de fouille de textes permettent d'extraire des connaissances inconnues d'un grand nombre d'articles, ce qui pourrait s'avérer indispensable pour spécifier l'organisation des construits liés à la somnolence (Gauld *et coll.*, 2020). La tentative présentée dans le chapitre 3 mériterait ainsi d'être explorée plus en profondeur.

Enfin, nous n'avons pas extrait d'informations concernant la fiabilité, la sensibilité au changement et les seuils pathologiques de ces outils de mesure de la somnolence, qui sont des paramètres essentiels pour leur utilisation clinique. Une analyse plus systématique de ces facteurs est nécessaire.

4.4 Conclusion

Cette revue générale est un premier pas vers la clarification du concept de somnolence. Même si cette première étape est essentielle pour mettre en évidence la représentation acceptée de la communauté de la recherche sur le sommeil au sens large concernant la somnolence, elle devrait encourager la communauté à mieux conceptualiser et modéliser, dans un modèle intégratif, les différentes composantes de la somnolence, façonnées par une approche computationnelle (Kriegeskorte et Douglas, 2018) comme cela est fait pour d'autres symptômes liés au cerveau (Friston *et coll.*, 2014). Cette clarification et cette modélisation pourraient être bénéfiques à la communauté, non seulement en contribuant à résoudre l'épineuse question des seuils médicaux pour le diagnostic et le traitement des troubles de l'hypersomnolence, mais aussi en permettant la spécification des mécanismes neurophysiologiques sous-jacents, dans le contexte des troubles du sommeil et de leur altération.

Chapitre 5

La somnolence et les construits qui y sont liés

Sommaire

5.1 Objectif du chapitre

Face aux précédentes tentatives infructueuses de faire émerger une structure relationnelle des différents construits autour de la somnolence à partir des articles publiés par la communauté de recherche en médecine du sommeil, nous proposons dans ce chapitre une organisation de la somnolence et des différents construits qui lui sont souvent associés (fatigue, perte de vigilance, *hypoarousal*). Cette mise en relation est le fruit de riches échanges avec des spécialistes en médecine du sommeil, dont Jean-Arthur Micoulaud-Franchi et Régis Lopez.

Nous proposons d'effectuer cette mise en relation selon deux granularités temporelles différentes : d'une part les états de somnolence au court terme et les construits psychophysiologiques "instantanés" avec lesquels ils sont en relation ; d'autre part les construits cliniques de l'hypersomnolence, qui se manifestent lorsque la somnolence devient un trait du sujet.

Tous les concepts développés dans ce chapitre sont représentés dans la figure 5.1, qui propose une version schématique des différents construits proposés et de leurs relations.

5.2 Somnolence à court terme – construits psychophysiologiques

L'échelle temporelle la plus fine que l'on peut définir de la somnolence est celle des processus neuropsychologiques en eux-mêmes, définissant notre perception de l'instantané (Bergson, 1889; Merchant *et coll.*, 2013).

En considérant des temporalités allant de l'instantané à des durées de l'ordre de la minute, voire de l'heure, les mécanismes psychophysiologiques liés à la somnolence sont organisés autour de trois grands pôles : l'*arousal* (qui pourrait se traduire par un « état résultant de l'activité des processus d'éveil [1] »), la somnolence, et les performances cognitives.

5.2.1 L'arousal

Dans le *Research Domain Criteria Initiative* – Rdoc – orchestré par le *National Institute of Mental Health* américain, l'*arousal* est défini comme « un continuum de sensibilité de l'organisme aux stimulus, externes comme internes [2] ».

L'arousal joue ici un rôle de filtre des stimulus : suivant leur intensité et le niveau d'arousal, ceux-ci seront – ou pas – pris en compte par les systèmes d'attention.

Ce mécanisme central peut être influencé par les facteurs suivants :

— Les rythmes circadiens et processus homéostatiques. Les phénomènes liés au rythme circadien – qui joue le rôle d'horloge interne – tels que la faim ou le sommeil influencent la sensibilité aux stimulus. Par exemple, à l'heure du sommeil, les processus circadiens et homéostatiques diminuent l'arousal afin de permettre la mise en sommeil.

— Le système d'éveil lui aussi a une influence sur l'arousal, dans le sens opposé des influences circadiennes et homéostatiques, en réaugmentant notre sensibilité aux stimuli lors de l'éveil.

Le niveau d'arousal peut être mesuré par exemple par des tests de potentiels évoqués [*Event Related Potential* en anglais, (Haider *et coll.*, 1964)].

1. *Le sommeil et ses pathologies : approche clinique transversale chez l'adulte et l'enfant*, Société française de recherche et médecine du sommeil, 2021 (Ellipses ed.)
2. https://www.nimh.nih.gov/research/research-funded-by-nimh/rdoc/definitions-of-the-rdoc-domains-and-constructs

5.2.2 La somnolence

Somnolence "objective" La somnolence est définie comme l'état de la personne sur un continuum allant du sommeil à un état d'éveil (au sens de la *wakefulness*). La somnolence physiologique est influencée par le niveau d'*arousal*, mais aussi par la quantité et la qualité de sommeil du sujet. Ces dernières sont principalement influencées par des facteurs comportementaux

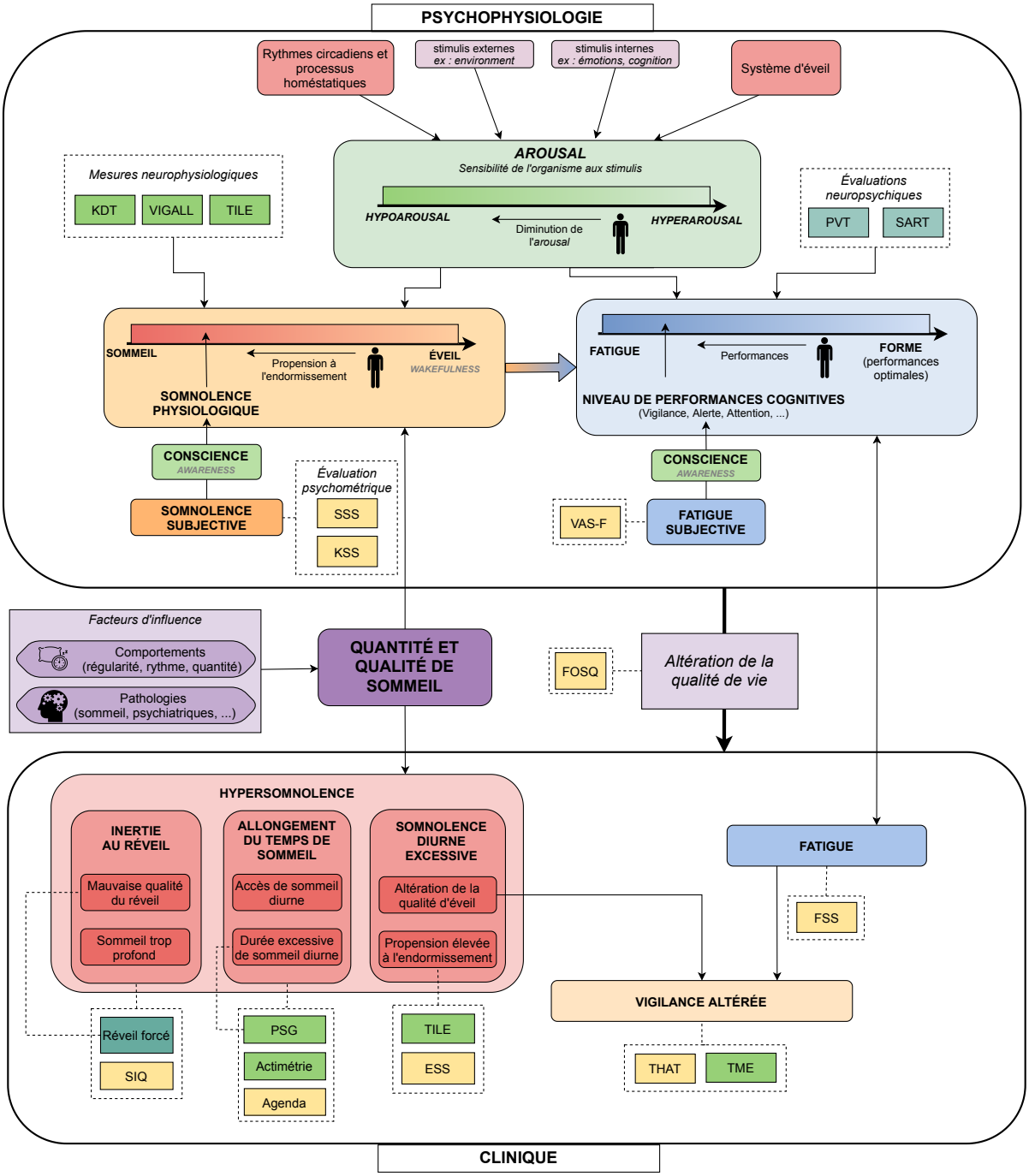

FIGURE 5.1 – Organisations psychophysiologiques et cliniques des concepts liés à la somnolence, et leurs mesures les plus courantes.

(régularité, rythmes et quantité de sommeil) et peuvent être altérées lorsque le sujet est atteint de pathologies du sommeil, mais aussi psychiatriques, ORL, cardiaques ...

La somnolence physiologique peut être mesurée par exemple grâce au test de vigilance de Karolinska (Åkerstedt et Gillberg, 1990, KDT) où à l'algorithme de vigilance de Leipzig (Huang *et coll.*, 2015, VIGALL), qui sont des mesures basées sur la puissance des signaux EEG dans certaines bandes de fréquences, mesurant le niveau de somnolence physiologique.

Propension à l'endormissement La vitesse à laquelle la somnolence physiologique se rapproche du sommeil s'appelle la propension à l'endormissement : c'est ce qui est mesuré par le TILE, décrit plus en détail dans le chapitre 7.

Somnolence subjective Un concept important lié au niveau de somnolence physiologique est celui de *somnolence subjective*. Celle-ci correspond à la conscience (au sens de l'*awareness*) que le sujet a de son niveau de somnolence physiologique. Cette opposition entre somnolence physiologique et somnolence subjective se retrouve dans les deux chapitres précédents, et est un axe majeur de la mesure de la somnolence.

Elle est mesurée par le biais de questionnaires psychométriques, qui peuvent prendre différentes formes. Les plus utilisées sont l'échelle de somnolence de Stanford (Hoddes *et coll.*, 1973, SSS) et l'échelle de somnolence de Karolinska (Åkerstedt et Gillberg, 1990, KSS) [3].

5.2.3 Le niveau de performances cognitives

Domaine à la fois riche et aux délimitations floues, nous proposons ici un découpage du concept de *niveau de performances cognitives* selon les dimensions que nous avons pu observer dans la littérature de la détection de la somnolence dans la voix. Ainsi, l'objet de cette partie est de proposer une description succincte des mécanismes associés à différents termes retrouvés dans la littérature, et d'appuyer leur différence avec la somnolence telle que nous l'avons définie précédemment.

Fatigue La distinction mérite d'autant plus d'être faite que dans le vocabulaire courant, les deux mots *fatigue* et *somnolence* sont employés couramment de manière interchangeable. Une liste des termes utilisés couramment comme synonymes de somnolence, fatigue, ou des deux en langue anglaise a été proposée dans (Hirshkowitz, 2013). Cette liste, reproduite dans le tableau 5.1, montre que la difficulté de décrire les construits en jeu apparaît dès leur dénomination, qui n'est pas clairement établie.

Dans ce document, en accord avec les définitions trouvées dans le chapitre 3, nous accepterons la définition suivante de la fatigue : « état d'épuisement des performances cognitives ». Sur un continuum allant de la fatigue à la forme, l'état dans lequel se trouve le sujet est son niveau de performances cognitives.

3. Il y a cependant une nuance à apporter sur la différence entre « mesure de la somnolence subjective » et « mesure subjective de la somnolence ». En effet, l'évaluation par un sujet de son propre niveau de somnolence se fait nécessairement par un processus de prise de conscience (là encore au sens de l'*awareness*), c'est-à-dire que l'outil de mesure de sa propre somnolence par un sujet est *nécessairement* son *awareness*. En toute rigueur, le terme « mesure subjective de la somnolence » a donc un sens flou. Il est cependant utilisé dans le vocabulaire courant de la clinique et de la recherche en médecine du sommeil, et sera utilisé comme un synonyme de « mesure de la somnolence subjective »

Fatigué	Somnolent	L'un ou les deux
Beat	*Crashing*	*Exhausted*
Languor	*Drowsy*	*Burned out*
Lassitude	*Fading*	*Bushed*
Lethargic	*Groggy*	*Gassed*
Listless	*Narcotized*	*Pooped*
Knackered	*Heavy-headed*	*Played-out*
Sluggish	*Punchy*	*Tired*
Weariness	*Gorked*	*Tuckered-out*
Whipped	*Yawny*	*Wiped*
Zoned	*Slap happy*	*Zonked*

TABLEAU 5.1 – Tableau extrait de (Hirshkowitz, 2013), présentant certains mots utilisés dans le vocabulaire anglais comme synonymes de somnolent (*sleepy*), fatigué (*fatigued*), ou des synonymes désignant l'un ou l'autre ou les deux de manière indistincte.

Fatigue subjective Contrairement à la somnolence, à laquelle on associe généralement un niveau d'activité physiologique, la fatigue ne peut être mesurée que de manière indirecte, par le biais de questionnaires psychométriques : ceux-ci ne permettant de mesurer que la conscience qu'a le sujet de son niveau de fatigue, et on parle ici de *fatigue subjective*.

Fatigue physique La fatigue physique, appelée parfois *exertion* en anglais, ne rentre pas dans le cadre des travaux présentés dans ce manuscrit. Elle n'a donc pas été représentée sur la figure 5.1.

Vigilance, alerte et attention De même, les concepts suivants sont présentés de manière brève, dans l'optique de les définir *par rapport* à la somnolence : la vigilance et l'alerte font l'objet à eux seuls de nombreux projets de recherche. Pour les définitions suivantes, nous nous appuyons sur la revue récente proposée dans (van Schie *et coll.*, 2021).

— Sans détailler outre mesure, nous réduisons ici l'*attention* à un processus par lequel la conscience du sujet se focalise sur un évènement extérieur ou intérieur. Chaque évènement agit comme une stimulation et rehausse le niveau d'attention jusqu'à ce que la stimulation cesse. L'attention soutenue est ainsi définie comme un état dans lequel un certain niveau d'attention est maintenu volontairement, afin par exemple d'augmenter ses capacités de réaction aux stimulus externes ;
— le *niveau d'alerte* est quant à lui défini comme la capacité d'un individu à mobiliser ses capacités d'attention ;
— la *vigilance* est définie comme la capacité à maintenir son attention au-dessus d'un certain seuil.

Lien entre *arousal* et *niveau d'alerte* Ces définitions étant posées, un lien subtil apparaît entre *niveau d'alerte* et *arousal* : alors que ce dernier définit la sensibilité de l'organisme aux stimulus, et différencie ce qui va être traité comme tel, le niveau d'alerte va définir la capacité à se focaliser sur ces stimulus. La réactivité à un stimuli externe dépend ainsi de la relation entre ces deux états : expliquer les performances cognitives sous le seul prisme de l'un ou de l'autre pris séparément est dénué de sens.

Le cas de la somnolence et de la fatigue au volant Au regard de ces définitions, il apparaît que ce qui est défini dans (Hu et Lodewijks, 2020) comme de la « fatigue active », caractérisée par un épuisement des ressources cognitives dû à un maintien prolongé de l'attention relève de la fatigue telle que définie dans cette section, alors que la « fatigue passive », caractérisée par une diminution de l'attention due à un manque de stimulation, est en réalité une diminution du niveau de *vigilance* du locuteur et non une forme de fatigue comme pourrait le laisser supposer le nom.

Le même exercice de formalisation pourrait être fait pour de nombreuses études qui n'étudie non pas la fatigue ou la somnolence, mais leur impact sur les performances cognitives (Shilov et Kashevnik, 2021; Satish *et coll.*, 2020; Sparrow *et coll.*, 2019; Poursadeghiyan *et coll.*, 2018; Vicente *et coll.*, 2016; Keelan et Mårtensson, 2017; Hu et Lodewijks, 2020; Vicente *et coll.*, 2016; Borghini *et coll.*, 2014; Sigari *et coll.*, 2013; Di Stasi *et coll.*, 2012).

Exemple :
Dans le cas de la conduite automobile, les performances de conduite et la vigilance peuvent être altérées à la fois par la fatigue et la somnolence. Lorsqu'une personne prend le volant en milieu de nuit, même pour un trajet court, sa somnolence due à sa pression de sommeil et à son faible niveau d'*arousal* altérera sa vigilance, même si le trajet n'est pas suffisamment long pour la fatiguer. Au contraire, une personne roulant en fin d'après-midi après avoir conduit de nombreuses heures consécutives sera fatigué par le maintien de son attention durant une trop longue période, et ses performances de conduite seront altérées sans nécessairement que son niveau de somnolence soit élevé.

5.3 Somnolence au long cours – construits cliniques

Alors qu'il est normal, au cours d'une période de 24h, de ressentir de la somnolence de manière ponctuelle (à l'heure du coucher par exemple), lorsque la somnolence représente un handicap pour les sujets et interfère avec leur fonctionnement quotidien, on parle alors de *somnolence excessive*. Les plaintes cliniques de somnolence sont souvent liées à des ressentis couvrant de longues périodes de temps : la somnolence devient peu à peu un trait des personnes en faisant l'expérience.

5.3.1 Composantes de l'hypersomnolence

Somnolence diurne excessive La composante la plus courante de l'hypersomnolence dans les plaintes cliniques concerne la somnolence diurne excessive (SDE), définie comme « la plainte d'une incapacité à rester éveillé sur les périodes d'éveil normal de la journée » (Lammers *et coll.*, 2020). Elle comprend à la fois des symptômes liés à une altération de la qualité d'éveil et une propension élevée à l'endormissement.

Cependant, la SDE n'est qu'une des trois composantes du syndrome d'hypersomnolence (Pertenais *et coll.*, 2019).

Quantité excessive de sommeil et inertie au réveil

Au cours des dernières décennies, l'étude des pathologies les plus sévères associées à une somnolence excessive (c'est-à-dire la narcolepsie et l'hypersomnie idiopathique) a conduit à l'émergence du concept d'hypersomnolence (Lammers *et coll.*, 2020). Ce syndrome associe une somnolence diurne excessive (qui fait surtout référence à une propension excessive au

sommeil et à la somnolence), à deux autres symptômes connexes : une quantité excessive de sommeil et une inertie du sommeil (Lammers *et coll.*, 2020).

Quantité excessive de sommeil La quantité de sommeil fait référence à la durée totale de sommeil obtenu pendant la période de sommeil principale ou sur 24h. La distinction entre quantité excessive de sommeil et long sommeil reste controversée : alors que le long sommeil serait défini sur la base de la durée du sommeil avec un seuil variant de >8 heures à >11 heures, la quantité excessive de sommeil serait liée à une détresse ou une altération significative en relation avec le long sommeil (Lammers *et coll.*, 2020).

Inertie du sommeil De plus, certains sujets éprouvent des difficultés à se réveiller après le sommeil et à « se mettre en route ». L'inertie du sommeil correspond à la forme la plus sévère de ce phénomène. L'inertie du sommeil se caractérise par une baisse de la vigilance et une altération des performances, pouvant durer jusqu'à plusieurs heures. Elle est parfois associée à une confusion, une difficulté à réagir de manière adéquate aux stimuli externes au réveil, dans un état souvent appelé *ivresse du sommeil* (Evangelista *et coll.*, 2021; Trotti *et coll.*, 2022).

Allongement du temps de sommeil La troisième composante de l'hypersomnolence – l'allongement du temps de sommeil – comprend deux symptômes liés à des durées excessives de sommeil, respectivement de jour et de nuit.

5.3.2 Le trouble d'hypersomnolence

Il est important de différencier l'hypersomnolence périphérique à une autre pathologie – de nombreuses pathologies comme l'apnée du sommeil ou la dépression peuvent être des sources de somnolence – de celle qui est centrale, et pour laquelle une pathologie du sommeil est sous-jacente. En effet, « la somnolence, la somnolence excessive, la somnolence diurne excessive, la somnolence subjective, la somnolence objective, l'hypersomnolence, l'hypersomnie, la capacité à s'endormir, l'incapacité à rester éveillé, l'hypovigilance, la fatigue et la fatigue excessive sont souvent employées pour définir cet état. » (Dauvilliers *et coll.*, 2017b).

À ces confusions s'ajoute le fait que les classifications proposées jusqu'au début des années 2010 ne facilitaient pas la clarification des concepts. C'est le cas par exemple de la deuxième version de l'*International classification of sleep disorders* – ISCD2 – qui mettait en avant deux définitions distinctes pour la somnolence excessive et la somnolence diurne. La première était définie comme la « somnolence inappropriée non désirée occurant une période durant laquelle on attend qu'il soit éveillé et alerte. »[4] tandis que la somnolence diurne est définie dans cette classification comme « l'incapacité à rester éveillé ou alerte durant la majeure partie des épisodes d'éveil de la journée, causant des accès de sommeil indésirables. »[5]

Un autre exemple est l'article publié par Kendzerska *et coll.* (2014), qui introduit trois nouveaux concepts : la propension à l'endormissement instantanée, la propension à l'endormissement situationnelle et la propension à l'endormissement moyenne[6]. Face à la profusion de construits et au nombre de combinaisons possibles (somnolence diurne/nocturne/les deux, somnolence dans une situation favorisant l'endormissement/stimulante/ni l'un ni l'autre ...)

4. "Excessive sleepiness is inappropriate or undesired sleepiness that occurs when an individual would usually be expected to be awake and alert."

5. "Daytime sleepiness is the inability to stay awake and alert during the major waking episodes of the day, resulting in unintended lapses into drowsiness and sleep"

6. respectivement *instantaneous sleep propensity, situational sleep propensity* et *average sleep propensity*

ce concept a été clarifié comme « trouble d'hypersomnolence » dans la cinquième édition du *Diagnostic and statistical manual of mental disorders* – DSM-5 – en 2013, puis dans la troisième version de l'ISCD en 2014.

5.3.3 Altération de la vigilance

En lien avec l'altération de la qualité d'éveil et la fatigue chronique, les sujets ayant des plaintes cliniques d'hypersomnolence peuvent présenter une altération de leur vigilance de façon chronique ou sur de longues périodes de temps. C'est ce qui est mesuré par le test d'alerte de Toronto (Shahid *et coll.*, 2016, THAT), qui mesure la plainte de baisse d'alerte quotidienne des sujets, ou par le TME, qui mesure la baisse de vigilance à travers la capacité à se maintenir éveillé en l'absence de stimulations.

5.4 Conclusion

Comme nous l'avons montré dans la précédente revue du chapitre 4, la définition de la *somnolence* et des différents termes qui lui sont liés est encore discutée et ne fait pas l'objet d'un consensus de la communauté. Nous avons donc fixé dans ce chapitre les définitions qui seront employées dans ce document, afin d'avoir un vocabulaire commun avec le lecteur et de faciliter sa compréhension. La communauté de recherche en médecine du sommeil gagnerait cependant à travailler à un consensus sur la question, ce qui faciliterait à la fois la communication entre chercheurs, mais permettrait aussi de penser un modèle relationnel et intégratif rigoureux de la somnolence.

Conclusion de la partie

Définition de la *somnolence*

Nous avons cherché dans cette partie à définir ce qu'est la « somnolence », d'abord à partir de trois dictionnaires de référence de la langue française. Ensuite, nous avons proposé deux approches de fouille de textes basées respectivement sur un nuage de mots et sur une analyse en réseau des termes retrouvés dans les titres, résumé et mots-clés de requête PubMed. Ces analyses ont permis d'avoir un aperçu des différentes oppositions qui structurent le construit de la somnolence :

— au court terme vs au long terme ;
— subjective vs objective ;
— normale vs pathologique ;
— en population générale vs en population pathologique ;
— somnolence centrale vs somnolence périphérique.

Revue générale des outils de mesure de la somnolence

Afin d'affiner les définitions précédentes, nous avons proposé une revue générale – basée sur 36 revues de la littérature – des différents outils de la somnolence. Les 99 outils identifiés ont été classés en 8 catégories par des médecins experts du sommeil, et pour chaque mesure, le premier article mentionnant celle-ci dans le contexte de la somnolence ou de la fatigue a été identifié. L'organisation de ces outils de mesures en fonction de leur catégorie et des années de publication des premiers articles en faisant mention reflète l'histoire de la médecine du sommeil, et témoigne d'une *dépendance au chemin*, entretenue par la formation, la pratique clinique, et l'utilisation des outils cliniques les plus validés – qui sont généralement les plus anciens. Cependant, cette approche n'a pas permis de distinguer la *somnolence* des autres construits qui lui sont proches – et parfois confondus dans la littérature. La direction de recherche la plus prometteuse pour arriver à un consensus sur la définition de la somnolence semble être l'utilisation plus approfondie des outils de fouille de textes, dont nous avons effectué une étude préliminaire dans le chapitre 3.

Proposition d'un modèle relationnel

Nous avons enfin proposé, à partir des définitions établies dans la littérature, des résultats des chapitres 3 et 4 et en collaboration avec des médecins spécialistes du sommeil, un schéma relationnel des différents construits liés à la somnolence et de leurs mesures les plus courantes, selon deux modalités temporelles (psychophysiologique "instantané" et clinique). Ce schéma permet de fixer les définitions que nous utiliserons dans la suite de ce manuscrit et de proposer un premier modèle d'interaction de ces construits.

Prochaine partie

Le choix d'une mesure de la somnolence pour annoter des données regroupées en base de données – notamment des enregistrements vocaux pour l'élaboration d'un corpus pour la

détection automatique de la somnolence – ne peut pas reposer uniquement sur des arguments médicaux, mais nécessite aussi une prise en compte des contraintes de traitement de données. Ces choix sont notamment exposés dans la prochaine partie, qui propose une étude approfondie des corpus existants sur le sujet de la détection automatique de la somnolence à partir de marqueurs vocaux, et qui dresse un ensemble de recommandations sur la constitution d'un tel corpus – et notamment quelle mesure choisir pour annoter les données.

Enfin, nous finirons cette partie sur une citation de Paul Feyerabend, dans *Against Method* (1975) :

> « without a constant misuse of language, there cannot be any discovery, any progress… »

Et si cette suspension de la définition de la somnolence était bénéfique à la communauté de recherche en médecine du sommeil ?

Bibliographie de la partie

Åkerstedt, T., Anund, A., Axelsson, J., et Kecklund, G. (**2014**). "Subjective sleepiness is a sensitive indicator of insufficient sleep and impaired waking function," Journal of sleep research **23**(3), 240–52.

Åkerstedt, T., et Gillberg, M. (**1990**). "Subjective and objective sleepiness in the active individual.," Int J Neurosci **52**, 29–37, doi: 10.3109/00207459008994241.

Anniss, A. M., Young, A., et O'Driscoll, D. M. (**2021**). "Microsleep assessment enhances interpretation of the Maintenance of Wakefulness Test," Journal of Clinical Sleep Medicine doi: 10.5664/jcsm.9250.

Arand, D., Bonnet, M., Hurwitz, T., Mitler, M., Rosa, R., et Sangal, R. B. (**2005**). "The Clinical Use of the MSLT and MWT," SLEEP **28**(1), 123–144, doi: 10.1093/sleep/28.1.123.

Bargiotas, P., Lachenmayer, M. L., Schreier, D. R., Mathis, J., et Bassetti, C. L. (**2019**). "Sleepiness and sleepiness perception in patients with Parkinson's disease : a clinical and electrophysiological study," Sleep **42**(4), zsz004, doi: 10.1093/sleep/zsz004.

Bastien, C. H., Vallières, A., et Morin, C. M. (**2001**). "Validation of the Insomnia Severity Index as an outcome measure for insomnia research," Sleep Medicine **2**(4), 297–307, doi: 10.1016/S1389-9457(00)00065-4.

Bennett, L., Stradling, J., et Davies, R. (**1997**). "A behavioural test to assess daytime sleepiness in obstructive sleep apnoea," Journal of Sleep Research **6**(2), 142–145, doi: 10.1046/j.1365-2869.1997.00039.x.

Bergson, H. (**1889**). *Essais sur les données immédiates de la conscience*, félix alcan éd.

Bioulac, S., Micoulaud-Franchi, J.-A., Arnaud, M., Sagaspe, P., Moore, N., Salvo, F., et Philip, P. (**2017**). "Risk of Motor Vehicle Accidents Related to Sleepiness at the Wheel : A Systematic Review and Meta-Analysis," Sleep **40**(10), doi: 10.1093/sleep/zsx134.

Bloom, P., et Keil, F. C. (**2001**). "Thinking through language," Mind & Language **16**(4), 351–367.

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., et Babiloni, F. (**2014**). "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," Neuroscience & Biobehavioral Reviews **44**, 58–75, doi: 10.1016/j.neubiorev.2012.10.003.

Brown, I. D., Simmonds, D. C. V., et Tickner, A. H. (**1967**). "Measurement of Control Skills, Vigilance, and Performance on a Subsidiary Task during 12 Hours of Car Driving," Ergonomics **10**(6), 665–673, doi: 10.1080/00140136708930920.

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., et Kupfer, D. J. (**1989**). "The Pittsburgh sleep quality index : A new instrument for psychiatric practice and research," Psychiatry Research **28**(2), 193–213, doi: 10.1016/0165-1781(89)90047-4.

Cai, A. W., Manousakis, J. E., Lo, T. Y., Horne, J. A., Howard, M. E., et Anderson, C. (**2021**). "I think I'm sleepy, therefore I am – Awareness of sleepiness while driving : A systematic review," Sleep Medicine Reviews **60**, 101533, doi: 10.1016/j.smrv.2021.101533.

Carskadon, M. A., et Dement, W. C. (**1979**). "Effects of Total Sleep Loss on Sleep Tendency," Perceptual and Motor Skills **48**(2), 495–506, doi: 10.2466/pms.1979.48.2.495.

Chaudhuri, K. R., Pal, S., DiMarco, A., Whately-Smith, C., Bridgman, K., Mathew, R., Pezzela, F. R., Forbes, A., Högl, B., et Trenkwalder, C. (**2002**). "The Parkinson's disease sleep scale : a new instrument for assessing sleep and nocturnal disability in Parkinson's disease," Journal of Neurology, Neurosurgery & Psychiatry **73**(6), 629–635, doi: 10.1136/jnnp.73.6.629.

Chu, J. S. G., et Evans, J. A. (**2021**). "Slowed canonical progress in large fields of science," Proceedings of the National Academy of Sciences **118**(41), doi: 10.1073/pnas.2021636118.

Crawford, A. (**1961**). "Fatigue and Driving," Ergonomics **4**(2), 143–154, doi: 10.1080/00140136108930515.

Daniel, R. S. (**1967**). "Alpha and Theta EEG in Vigilance," Perceptual and Motor Skills **25**(3), 697–703, doi: 10.2466/pms.1967.25.3.697.

Dauvilliers, Y., Arnulf, I., Foldvary-Schaefer, N., Morse, A. M., Šonka, K., Thorpy, M. J., Mignot, E., Chandler, P., Parvataneni, R., Black, J., Sterkel, A., Chen, D., Skobieranda, F., et Bogan, R. K. (**2022**). "Safety and efficacy of lower-sodium oxybate in adults with idiopathic hypersomnia : a phase 3, placebo-controlled, double-blind, randomised withdrawal study," The Lancet Neurology **21**(1), 53–65, doi: 10.1016/S1474-4422(21)00368-9.

Dauvilliers, Y., Evangelista, E., Barateau, L., Lopez, R., Chenini, S., Delbos, C., Beziat, S., et Jaussent, I. (**2019**). "Measurement of symptoms in idiopathic hypersomnia : The Idiopathic Hypersomnia Severity Scale," Neurology **92**(15), e1754–e1762, doi: 10.1212/WNL.0000000000007264.

Dauvilliers, Y., Evangelista, E., de Verbizier, D., Barateau, L., et Peigneux, P. (**2017**a). "[18F]Fludeoxyglucose-Positron Emission Tomography Evidence for Cerebral Hypermetabolism in the Awake State in Narcolepsy and Idiopathic Hypersomnia," Frontiers in Neurology **8**, 350, doi: 10.3389/fneur.2017.00350.

Dauvilliers, Y., Lopez, R., et Lecendreux, M. (**2017**b). "Consensus. Hypersomnolence : évaluation et limites nosographiques," Médecine du Sommeil **14**(3), 132–137, doi: 10.1016/j.msom.2017.07.004.

Davies, D. R., et Krkovic, A. (**1965**). "Skin-Conductance, Alpha-Activity, and Vigilance," The American Journal of Psychology **78**(2), 304, doi: 10.2307/1420507.

Des Champs de Boishebert, L., Pradat, P., Bastuji, H., Ricordeau, F., Gormand, F., Le Cam, P., Stauffer, E., Petitjean, T., et Peter-Derex, L. (**2021**). "Microsleep versus Sleep Onset Latency during Maintenance Wakefulness Tests : Which One Is the Best Marker of Sleepiness ?," Clocks & Sleep **3**(2), 259–273, doi: 10.3390/clockssleep3020016.

Di Stasi, L. L., Renner, R., Catena, A., Cañas, J. J., Velichkovsky, B. M., et Pannasch, S. (**2012**). "Towards a driver fatigue test based on the saccadic main sequence : A partial validation by subjective report data," Transportation Research Part C : Emerging Technologies **21**(1), 122–133, doi: 10.1016/j.trc.2011.07.002.

Dinges, D. F., et Powell, J. W. (**1985**). "Microcomputer analyses of performance on a portable, simple visual RT task during sustained operations," Behavior Research Methods, Instruments, & Computers **17**(6), 652–655, doi: 10.3758/BF03200977.

Drakatos, P., Suri, A., Higgins, S. E., Ebrahim, I. O., Muza, R. T., Kosky, C. A., Williams, A. J., et Leschziner, G. D. (**2013**). "Sleep stage sequence analysis of sleep onset REM periods in the hypersomnias," Journal of Neurology, Neurosurgery & Psychiatry **84**(2), 223–227, doi: 10.1136/jnnp-2012-303578.

Esfandyarpour, R., Kashi, A., Nemat-Gorgani, M., Wilhelmy, J., et Davis, R. W. (**2019**). "A nanoelectronics-blood-based diagnostic biomarker for myalgic encephalomyelitis/chronic fatigue syndrome (ME/CFS)," Proceedings of the National Academy of Sciences **116**(21), 10250–10257, doi: 10.1073/pnas.1901274116.

Evangelista, E., Lopez, R., Barateau, L., Chenini, S., Bosco, A., Jaussent, I., et Dauvilliers, Y. (**2018**). "Alternative diagnostic criteria for idiopathic hypersomnia : A 32-hour protocol," Annals of Neurology **83**(2), 235–247, doi: 10.1002/ana.25141.

Evangelista, E., Rassu, A. L., Lopez, R., Biagioli, N., Chenini, S., Barateau, L., Jaussent, I., et Dauvilliers, Y. (**2021**). "Sleep inertia measurement with the psychomotor vigilance task in idiopathic hypersomnia," Sleep **zsab220**, doi: 10.1093/sleep/zsab220.

Fernandez-Mendoza, J., Puzino, K., Amatrudo, G., Bourchtein, E., Calhoun, S. L., Plante, D. T., et Kaplan, K. (**2021**). "The Hypersomnia Severity Index : reliability, construct, and criterion validity in a clinical sample of patients with sleep disorders," Journal of Clinical Sleep Medicine doi: 10.5664/jcsm.9426.

Friston, K. J., Stephan, K. E., Montague, R., et Dolan, R. J. (**2014**). "Computational psychiatry : the brain as a phantastic organ," The Lancet Psychiatry **1**(2), 148–158, doi: 10.1016/S2215-0366(14)70275-5.

Gauld, C., et Micoulaud-Franchi, J. A. (**2021**a). "Analyse en réseau par fouille de données textuelles systématique du concept de psychiatrie personnalisée et de précision," L'Encéphale **47**(4), 341–347, doi: 10.1016/j.encep.2020.08.008.

Gauld, C., et Micoulaud-Franchi, J.-A. (**2021**b). "Why could sleep medicine never do without polysomnography ?," Journal of Sleep Research doi: 10.1111/jsr.13541.

Gauld, C., Ouazzani, K., et Micoulaud-Franchi, J.-A. (**2020**). "Commentary on Lammers et al. "Diagnosis of central disorders of hypersomnolence : A reappraisal by European experts" : From clinic to clinic via ontology and semantic analysis on a bullet point path," Sleep Medicine Reviews **52**, 101328, doi: 10.1016/j.smrv.2020.101328.

Giere, R. N. (**2010**). *Scientific perspectivism* (University of Chicago press).

Gool, J. K., Cross, N., Fronczek, R., Lammers, G. J., van der Werf, Y. D., et Dang-Vu, T. T. (**2020**). "Neuroimaging in Narcolepsy and Idiopathic Hypersomnia : from Neural Correlates to Clinical Practice," Current Sleep Medicine Reports **6**(4), 251–266, doi: 10.1007/s40675-020-00185-9.

Grant, M. J., et Booth, A. (**2009**). "A typology of reviews : an analysis of 14 review types and associated methodologies," Health Information & Libraries Journal **26**(2), 91–108, doi: 10.1111/j.1471-1842.2009.00848.x.

Guaita, M., Salamero, M., Vilaseca, I., Iranzo, A., Montserrat, J. M., Gaig, C., Embid, C., Romero, M., Serradell, M., León, C., de Pablo, J., et Santamaria, J. (**2015**). "The Barcelona Sleepiness Index : A New Instrument to Assess Excessive Daytime Sleepiness in Sleep Disordered Breathing," Journal of clinical sleep medicine **11**(11), 1289–1298, doi: 10.5664/jcsm.5188.

Haider, M., Spong, P., et Lindsley, D. B. (**1964**). "Attention, Vigilance, and Cortical Evoked-Potentials in Humans," Science **145**(3628), 180–182, doi: 10.1126/science.145.3628.180.

Heimstra, N. W. (**1970**). "The Effects of 'Stress Fatigue' on Performance in a Simulated Driving Situation," Ergonomics **13**(2), 209–218, doi: 10.1080/00140137008931134.

Hempel, C. G. (**1970**). "Aspects of Scientific Explanation and Other Essays in the Philosophy of Science," Philosophy of Science **37**(2), 312–314, doi: 10.1086/288305.

Hermans, M. C., Merkies, I. S., Laberge, L., Blom, E. W., Tennant, A., et Faber, C. G. (**2013**). "Fatigue and daytime sleepiness scale in myotonic dystrophy type 1," Muscle & Nerve **47**(1), 89–95, doi: 10.1002/mus.23478.

Herscovitch, J., et Broughton, R. (**1981**). "Sensitivity of the Stanford Sleepiness Scale to the Effects of Cumulative Partial Sleep Deprivation and Recovery Oversleeping," Sleep **4**(1), 83–92, doi: 10.1093/sleep/4.1.83.

Hertig-Godeschalk, A., Skorucak, J., Malafeev, A., Achermann, P., Mathis, J., et Schreier, D. R. (**2020**). "Microsleep episodes in the borderland between wakefulness and sleep," Sleep **43**(1), zsz163, doi: 10.1093/sleep/zsz163.

Hirshkowitz, M. (**2013**). "Fatigue, Sleepiness, and Safety," Sleep Medicine Clinics **8**(2), 183–189, doi: 10.1016/j.jsmc.2013.04.001.

Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., et Dement, W. C. (**1973**). "Quantification of Sleepiness : A New Approach," Psychophysiology **10**(4), 431–436, doi: 10.1111/j.1469-8986.1973.tb00801.x.

Honda, T., Fujiyama, T., Miyoshi, C., Ikkyu, A., Hotta-Hirashima, N., Kanno, S., Mizuno, S., Sugiyama, F., Takahashi, S., Funato, H., et Yanagisawa, M. (**2018**). "A single phosphorylation site of SIK3 regulates daily sleep amounts and sleep need in mice," Proceedings of the National Academy of Sciences **115**(41), 10458–10463, doi: 10.1073/pnas.1810823115.

Hu, X., et Lodewijks, G. (**2020**). "Detecting fatigue in car drivers and aircraft pilots by using non-invasive measures : The value of differentiation of sleepiness and mental fatigue," Journal of Safety Research **72**, 173–187, doi: 10.1016/j.jsr.2019.12.015.

Huang, J., Sander, C., Jawinski, P., Ulke, C., Spada, J., Hegerl, U., et Hensch, T. (**2015**). "Test-retest reliability of brain arousal regulation as assessed with VIGALL 2.0," Neuropsychiatric Electrophysiology **1**(1), 13, doi: 10.1186/s40810-015-0013-9.

Ingravallo, F., Vignatelli, L., Pagotto, U., Vandi, S., Moresco, M., Mangiaruga, A., Oriolo, C., Zenesini, C., Pizza, F., et Plazzi, G. (**2020**). "Protocols of a diagnostic study and a randomized controlled non-inferiority trial comparing televisits vs standard in-person outpatient visits for narcolepsy diagnosis and care : TElemedicine for NARcolepsy (TENAR)," BMC Neurology **20**(1), 176, doi: 10.1186/s12883-020-01762-9.

Johns, M. W. (**1991**). "A New Method for Measuring Daytime Sleepiness : The Epworth Sleepiness Scale," Sleep **14**(6), 540–545, doi: https://doi.org/10.1093/sleep/14.6.540.

Kaida, K., Takahashi, M., Åkerstedt, T., Nakata, A., Otsuka, Y., Haratani, T., et Fukasawa, K. (**2006**). "Validation of the Karolinska sleepiness scale against performance and EEG variables," Clinical Neurophysiology **117**(7), 1574–1581, doi: 10.1016/j.clinph.2006.03.011.

Kanady, J. C., et Harvey, A. G. (**2015**). "Development and Validation of the Sleep Inertia Questionnaire (SIQ) and Assessment of Sleep Inertia in Analogue and Clinical Depression," Cognitive Therapy and Research **39**(5), 601–612, doi: 10.1007/s10608-015-9686-4.

Kaplan, K. A., Plante, D. T., Cook, J. D., et Harvey, A. G. (**2019**). "Development and validation of the Hypersomnia Severity Index (HSI) : A measure to assess hypersomnia severity and impairment in psychiatric disorders," Psychiatry Research **281**, 112547, doi: 10.1016/j.psychres.2019.112547.

Kawai, R., Watanabe, A., Fujita, S., Hirose, M., Esaki, Y., Arakawa, C., Iwata, N., et Kitajima, T. (**2020**). "Utility of the sleep stage sequence preceding sleep onset REM periods for the diagnosis of narcolepsy : a study in a Japanese cohort," Sleep Medicine **68**, 9–17, doi: 10.1016/j.sleep.2019.04.008.

Keelan, O., et Mårtensson, H. (**2017**). *Feature Engineering and Machine Learning for Driver Sleepiness Detection*.

Kendzerska, T. B., Smith, P. M., Brignardello-Petersen, R., Leung, R. S., et Tomlinson, G. A. (**2014**). "Evaluation of the measurement properties of the Epworth sleepiness scale : a systematic review," Sleep Medicine Reviews **18**(4), 321–331, doi: 10.1016/j.smrv.2013.08.002.

Krahn, L. E., Arand, D. L., Avidan, A. Y., Davila, D. G., DeBassio, W. A., Ruoff, C. M., et Harrod, C. G. (**2021**). "Recommended protocols for the Multiple Sleep Latency Test and Maintenance of Wakefulness Test in adults : guidance from the American Academy of Sleep Medicine," Journal of Clinical Sleep Medicine **17**(12), 2489–2498, doi: 10.5664/jcsm.9620.

Kratzel, L., Glos, M., Veauthier, C., Rekow, S., François, C., Fietze, I., et Penzel, T. (**2021**). "Video-based sleep detection using ocular signals under the standard conditions of the maintenance of wakefulness test in patients with sleep disorders," Physiological Measurement **42**(1), 014004, doi: 10.1088/1361-6579/abdb7e.

Kriegeskorte, N., et Douglas, P. K. (**2018**). "Cognitive computational neuroscience," Nature neuroscience **21**(9), 1148–1160.

Lammers, G. J., Bassetti, C. L., Dolenc-Groselj, L., Jennum, P. J., Kallweit, U., Khatami, R., Lecendreux, M., Manconi, M., Mayer, G., Partinen, M., Plazzi, G., Reading, P. J., Santamaria, J., Sonka, K., et Dauvilliers, Y. (**2020**). "Diagnosis of central disorders of hypersomnolence : A reappraisal by European experts," Sleep Medicine Reviews **52**, 101306, doi: 10.1016/j.smrv.2020.101306.

Liebowitz, S., et Margolis, S. E. (**2014**). "Path Dependence and Lock-In," Books.

Lim, D. C., Mazzotti, D. R., Sutherland, K., Mindel, J. W., Kim, J., Cistulli, P. A., Magalang, U. J., Pack, A. I., de Chazal, P., et Penzel, T. (**2020**). "Reinventing polysomnography in the age of precision medicine," Sleep Medicine Reviews **52**, 101313, doi: 10.1016/j.smrv.2020.101313.

Littner, M. R., Kushida, C., Wise, M., Davila, D. G., Morgenthaler, T., Lee-Chiong, T., Hirshkowitz, M., Loube, D. L., Bailey, D., Berry, R. B., Kapen, S., et Kramer, M. (**2005**). "Practice Parameters for Clinical Use of the Multiple Sleep Latency Test and the Maintenance of Wakefulness Test," Sleep **28**(1), 113–121, doi: 10.1093/sleep/28.1.113.

Lowenstein, O., Feinberg, R., et Loewenfeld, I. E. (**1963**). *Pupillary movements during acute and chronic fatigue : A new test for the objective evaluation of tiredness*, **65** (Federal Aviation Agency, Office of Aviation Medicine).

Merchant, H., Harrington, D. L., et Meck, W. H. (**2013**). "Neural Basis of the Perception and Estimation of Time," Annual Review of Neuroscience **36**(1), 313–336, doi: 10.1146/annurev-neuro-062012-170349.

Mitler, M. M., Gujavarty, K. S., et Browman, C. P. (**1982**). "Maintenance of wakefulness test : A polysomnographic technique for evaluating treatment efficacy in patients with excessive somnolence," Electroencephalography and Clinical Neurophysiology **53**(6), 658–661, doi: 10.1016/0013-4694(82)90142-0.

Miyagawa, T., Miyadera, H., Tanaka, S., Kawashima, M., Shimada, M., Honda, Y., Tokunaga, K., et Honda, M. (**2011**). "Abnormally Low Serum Acylcarnitine Levels in Narcolepsy Patients," Sleep **34**(3), 349–353, doi: 10.1093/sleep/34.3.349.

Morrone, E., D'Artavilla Lupo, N., Trentin, R., Pizza, F., Risi, I., Arcovio, S., et Fanfulla, F. (**2020**). "Microsleep as a marker of sleepiness in obstructive sleep apnea patients," Journal of Sleep Research **29**(2), e12882, doi: 10.1111/jsr.12882.

Murer, T., Imbach, L. L., Hackius, M., Taddei, R. N., Werth, E., Poryazova, R., Gavrilov, Y. V., Winkler, S., Waldvogel, D., Baumann, C. R., et Valko, P. O. (**2017**). "Optimizing MSLT Specificity in Narcolepsy With Cataplexy," Sleep **40**(12), zsx173, doi: 10.1093/sleep/zsx173.

Onen, F., Lalanne, C., Pak, V. M., Gooneratne, N., Falissard, B., et Onen, S.-H. (**2016**). "A Three-Item Instrument for Measuring Daytime Sleepiness : The Observation and Interview Based Diurnal Sleepiness Inventory (ODSI)," J Clin Sleep Med. **12**(4), 505–512.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., McGuinness, L. A., Stewart, L. A., Thomas, J., Tricco, A. C., Welch, V. A., Whiting, P., et Moher, D. (**2021**). "The PRISMA 2020 statement : An updated guideline for reporting systematic reviews," PLoS medicine **18**(3), e1003583, doi: 10.1371/journal.pmed.1003583.

Pajcin, M., Banks, S., White, J. M., Dorrian, J., Paech, G. M., Grant, C., Johnson, K., Tooley, K., Fidock, J., Kamimori, G. H., et Della Vedova, C. B. (**2017**). "Decreased salivary alpha-amylase levels are associated with performance deficits during sleep loss," Psychoneuroendocrinology **78**, 131–141, doi: 10.1016/j.psyneuen.2017.01.028.

Pertenais, C., Lopez, R., Guichard, K., Dauvilliers, Y., Philip, P., Jaussent, I., et Micoulaud-Franchi, J.-A. (**2019**). "Revue de la littérature des outils psychométriques d'évaluation de la somnolence, de l'hypersomnolence et des hypersomnies chez l'adulte," Médecine du Sommeil **16**(4), 238–253, doi: 10.1016/j.msom.2019.08.001.

Pickering, A. (**1993**). "The Mangle of Practice : Agency and Emergence in the Sociology of Science," American Journal of Sociology **99**(3), 559–589, doi: 10.1086/230316.

Pizza, F., Barateau, L., Jaussent, I., Vandi, S., Antelmi, E., Mignot, E., Dauvilliers, Y., Plazzi, G., et Group, f. t. M. S. (**2019**). "Validation of Multiple Sleep Latency Test for the diagnosis of pediatric narcolepsy type 1," Neurology **93**(11), e1034–e1044, doi: 10.1212/WNL.0000000000008094.

Poursadeghiyan, M., Mazloumi, A., Nasl Saraji, G., Baneshi, M. M., Khammar, A., et Ebrahimi, M. H. (**2018**). "Using Image Processing in the Proposed Drowsiness Detection System Design," Iranian Journal of Public Health **47**(9), 1371–1378.

Radder, H. (**2006**). *The world observed/the world conceived* (University of Pittsburgh Press).

Rassu, A. L., Evangelista, E., Barateau, L., Chenini, S., Lopez, R., Jaussent, I., et Dauvilliers, Y. (**2021**). "Idiopathic Hypersomnia Severity Scale to better quantify symptoms severity and their consequences in idiopathic hypersomnia," Journal of Clinical Sleep Medicine doi: 10.5664/jcsm.9682.

Robertson, I. H., Manly, T., Andrade, J., Baddeley, B. T., et Yiend, J. (**1997**). "Oops!' : performance correlates of everyday attentional failures in traumatic brain injured and normal subjects," Neuropsychologia **35**(6), 747–758.

Rosenthal, L., Roehrs, T. A., et Roth, T. (**1993**). "The sleep-wake activity inventory : A self-report measure of daytime sleepiness," Biological Psychiatry **34**(11), 810–820, doi: https://doi.org/10.1016/0006-3223(93)90070-T.

Sagaspe, P., Micoulaud-Franchi, J.-A., Bioulac, S., Taillard, J., Guichard, K., Bonhomme, E., Dauvilliers, Y., Bastien, C. H., et Philip, P. (**2021**). "Self-perceived sleep during the Maintenance of Wakefulness Test : how does it predict accidental risk in patients with sleep disorders ?," Sleep **44**(11), zsab159, doi: 10.1093/sleep/zsab159.

Satish, K., Lalitesh, A., Bhargavi, K., Prem, M., et Anjali., T. (**2020**). "Driver Drowsiness Detection," dans *2020 International Conference on Communication and Signal Processing (ICCSP)*, IEEE, Chennai, India, pp. 0380–0384, doi: 10.1109/ICCSP48568.2020.9182237.

Shahid, A., Chung, S., Maresky, L., Danish, A., Bingeliene, A., Shen, J., et Shapiro, C. (**2016**). "The Toronto Hospital Alertness Test scale : relationship to daytime sleepiness, fatigue, and symptoms of depression and anxiety," Nature and Science of Sleep 41, doi: 10.2147/NSS.S91928.

Shen, J., Barbera, J., et Shapiro, C. M. (**2006**). "Distinguishing sleepiness and fatigue : focus on definition and measurement," Sleep Medicine Reviews **10**(1), 63–76, doi: 10.1016/j.smrv.2005.05.004.

Shilov, N., et Kashevnik, A. (**2021**). "An Effort to Detect Vehicle Driver's Drowsy State Based on the Speed Analysis," dans *2021 29th Conference of Open Innovations Association (FRUCT)*, pp. 324–329, doi: 10.23919/FRUCT52173.2021.9435466.

Sigari, M.-H., Fathy, M., et Soryani, M. (**2013**). "A Driver Face Monitoring System for Fatigue and Distraction Detection," International Journal of Vehicular Technology **2013**, 1–11, doi: 10.1155/2013/263983.

Smith, J. M., et Smith, D. C. P. (**1977**). "Database abstractions : aggregation and generalization," ACM Transactions on Database Systems **2**(2), 105–133, doi: 10.1145/320544.320546.

Sparrow, A. R., LaJambe, C. M., et Van Dongen, H. P. (**2019**). "Drowsiness measures for commercial motor vehicle operations," Accident Analysis & Prevention **126**, 146–159.

Tanida, K., Shimada, M., Khor, S.-S., Toyoda, H., Kato, K., Kotorii, N., Kotorii, T., Ariyoshi, Y., Kato, T., Hiejima, H., Ozone, M., Uchimura, N., Ikegami, A., Kume, K., Kanbayashi, T., Imanishi, A., Kamei, Y., Hida, A., Wada, Y., Kuroda, K., Miyamoto, M., Hirata, K., Takami, M., Yamada, N., Okawa, M., Omata, N., Kondo, H., Kodama, T., Inoue, Y., Mishima, K., Honda, M., Tokunaga, K., et Miyagawa, T. (**2021**). "Genome-wide association study of idiopathic hypersomnia in a Japanese population," Sleep and Biological Rhythms doi: 10.1007/s41105-021-00349-2.

Trotti, L. M., Saini, P., Bremer, E., Mariano, C., Moron, D., Rye, D. B., et Bliwise, D. L. (**2022**). "The Psychomotor Vigilance Test as a measure of alertness and sleep inertia in people with central disorders of hypersomnolence," Journal of Clinical Sleep Medicine jcsm.9884, doi: 10.5664/jcsm.9884.

Valko, P. O., Hunziker, S., Graf, K., Werth, E., et Baumann, C. R. (**2021**). "Sleep-wake misperception. A comprehensive analysis of a large sleep lab cohort," Sleep Medicine **88**, 96–103, doi: 10.1016/j.sleep.2021.10.023.

Vallat, R., Meunier, D., Nicolas, A., et Ruby, P. (**2019**). "Hard to wake up ? The cerebral correlates of sleep inertia assessed using combined behavioral, EEG and fMRI measures," NeuroImage **184**, 266–278, doi: 10.1016/j.neuroimage.2018.09.033.

van Schie, M. K., Lammers, G. J., Fronczek, R., Middelkoop, H. A., et van Dijk, J. G. (**2021**). "Vigilance : discussion of related concepts and proposal for a definition," Sleep Medicine **83**, 175–181, doi: 10.1016/j.sleep.2021.04.038.

Vicente, J., Laguna, P., Bartra, A., et Bailón, R. (**2016**). "Drowsiness detection using heart rate variability," Medical & Biological Engineering & Computing **54**(6), 927–937, doi: 10.1007/s11517-015-1448-7.

Violani, C., Lucidi, F., Robusto, E., Devoto, A., Zucconi, M., et Strambi, L. F. (**2003**). "The assessment of daytime sleep propensity : a comparison between the Epworth Sleepiness Scale and a newly developed Resistance to Sleepiness Scale," Clinical Neurophysiology **114**(6), 1027–1033, doi: https://doi.org/10.1016/S1388-2457(03)00061-0.

Wang, H., Lane, J. M., Jones, S. E., Dashti, H. S., Ollila, H. M., Wood, A. R., van Hees, V. T., Brumpton, B., Winsvold, B. S., Kantojärvi, K., Palviainen, T., Cade, B. E., Sofer, T., Song, Y., Patel, K., Anderson, S. G., Bechtold, D. A., Bowden, J., Emsley, R., Kyle, S. D., Little, M. A., Loudon, A. S., Scheer, F. A. J. L., Purcell, S. M., Richmond, R. C., Spiegelhalder, K., Tyrrell, J., Zhu, X., Hublin, C., Kaprio, J. A., Kristiansson, K., Sulkava, S., Paunio, T., Hveem, K., Nielsen, J. B., Willer, C. J., Zwart, J.-A., Strand, L. B., Frayling, T. M., Ray, D., Lawlor, D. A., Rutter, M. K., Weedon, M. N., Redline, S., et Saxena, R. (**2019**). "Genome-wide association analysis of self-reported daytime sleepiness identifies 42 loci that suggest biological subtypes," Nature Communications **10**(1), 3503, doi: 10.1038/s41467-019-11456-7.

Weaver, T. E., Laizner, A. M., Evans, L. K., Maislin, G., Chugh, D. K., Lyon, K., Smith, P. L., Schwartz, A. R., Redline, S., Pack, A. I., et others (**1997**). "An instrument to measure functional status outcomes for disorders of excessive sleepiness," Sleep **20**(10), 835–843.

Troisième partie

Corpus pour la détection de la somnolence dans la voix

Résumé

La détection automatique de la somnolence grâce à des marqueurs vocaux est un sujet relativement peu étudié comparé à d'autres pathologies ou symptômes.

À notre connaissance, il existe seulement quatre corpus ayant une taille suffisante pour la conception de systèmes d'apprentissage automatique.

Les deux premiers – le *Sleepy Language Corpus* (SLC) et le corpus SLEEP – ont été introduits à la communauté respectivement lors de compétitions proposées pour Interspeech 2011 et 2019. Ceux-ci contiennent des échantillons vocaux de sujets sains, annotés avec l'échelle de somnolence de Karolinska. Les deux derniers – le corpus TME et la base TILE – ont eux été enregistrés au CHU de Bordeaux, en étroite collaboration avec la laboratoire SANPSY. Les corpus SLC, SLEEP et la base TME sont exhaustivement présentés dans le chapitre 6 tandis que la base TILE, représentant une contribution majeure des travaux rapportés par ce document, est présentée dans le chapitre 7.

Nous proposons ensuite de comparer les caractéristiques de ces quatre corpus dans le chapitre 8, afin de faire émerger des recommandations sur la façon de construire un tel corpus, présentées dans le chapitre 9.

Enfin, nous proposons dans le chapitre 10 une première étude perceptuelle menée sur la base TILE, investiguant la faisabilité de la tâche de détection de la somnolence dans la voix par l'oreille humaine, étude préliminaire à sa faisabilité par des algorithmes.

Mots-clés

Corpus de l'état de l'art ; Méthodologie de conception de corpus ; Étude perceptuelle

Publications associées

Martin, V. P., Rouas, J.-L., Micoulaud-Franchi, J.-A., et Philip, P. (2020). The Objective and Subjective Sleepiness Voice Corpora. *12th Language Resources and Evaluation Conference*, 6525-6533. https://aclanthology.org/2020.lrec-1.803

Martin, V. P., Rouas, J.-L., Micoulaud-Franchi, J.-A., Philip, P., et Krajewski, J. (2021). How to Design a Relevant Corpus for Sleepiness Detection Through Voice ? *Frontiers in Digital Health*, 3, 124. https://doi.org/10.3389/fdgth.2021.686068

Martin, V. P., Ferron, A., Rouas, J.-L., Shochi, T., Philip, P., et Dupuy, L. (2022). Is human hearing able to estimate sleepiness from voice ? [En cours d'évaluation par les pairs].

Martin, V. P., Ferron, A., Rouas, J.-L., et Philip, P. (2022). "Prediction of Sleepiness Ratings from Voice by Man and Machine" : the Endymion replication study [En cours d'évaluation par les pairs].

Chapitre 6

Corpus de l'état de l'art pour la détection automatique de la somnolence

Sommaire

Ce chapitre présente les caractéristiques des corpus SLC, SLEEP et de la base TME selon les trois axes suivants :

1. Population et tâche vocale : sur quelle population a été enregistré le corpus, et quelles sont les tâches sur lesquelles les sujets ont été enregistrés ?
2. Annotation de la somnolence : comment la somnolence a-t-elle été annotée dans le corpus ?
3. Métadonnées : quelles sont les informations supplémentaires disponibles sur les locuteurs ou échantillons ?

6.1 *Sleepy Language Corpus* – SLC

Enregistré à l'institut de psychophysiologie de Düsseldorf et à l'institut des technologies de la sécurité de l'Université de Wuppertal, le *Sleepy Language Corpus* – SLC – a été introduit lors du challenge sur la détection d'état du locuteur proposé lors de la conférence internationale Interspeech 2011. Ce corpus inclut 9089 échantillons, produits par 99 locuteurs différents. Avant la diffusion du corpus SLEEP en 2019, le SLC a été le corpus international de référence pour tous les systèmes états-de-l'art (Cummins *et coll.*, 2018). La division en sous-corpus d'entraînement, de développement et de test est fournie dans le corpus, et lors de la compétition, l'annotation du sous-corpus de test était inconnue des compétiteurs. Dans cette partie, qui vise à décrire le contenu des corpus indépendamment des techniques d'apprentissage automatique qui seront ensuite appliquées, nous ne tiendrons pas compte de cette division.

6.1.1 Population et tâches vocales

Le SLC est l'agrégation de six études portant sur des privations partielles ou totales de sommeil sur des sujets sains (Schuller *et coll.*, 2013). Les sujets sont des volontaires issus de la population générale dont les éventuels problèmes de sommeil sont filtrés grâce à l'index de qualité du sommeil de Pittsburgh – PSQI (Buysse *et coll.*, 1989).

Les échantillons sont enregistrés soit dans un simulateur de conduite (décrit avec précision dans (Golz *et coll.*, 2007)), soit dans une salle de lecture. Les tâches vocales effectuées par les sujets sont réparties en cinq catégories :

— La lecture de quatre simulations de communication avec une tour de contrôle de trafic aérien, en anglais – noté *Avion* ;
— La lecture de *Die Sonne und der Nordwind*, la version en allemand de la fable «La bise et le soleil» – noté *Fable* ;
— La lecture de simulations de commande pour un système d'assistance à la conduite – noté *Conduite* ;
— Une autoprésentation du locuteur (parole spontanée) ;
— Des voyelles soutenues, « naturelles », « fortes » et en souriant.

Excepté la lecture des simulations d'interaction avec une tour de contrôle aérien qui sont en anglais, tous les autres échantillons sont en allemand. Le nombre d'échantillons enregistrés au cours de chaque tâche est représenté dans la figure 6.1.

6.1.2 Annotation de la somnolence – KSS

Dans le SLC, les enregistrements sont annotés avec la version allemande de la version à 10 niveaux de l'échelle de somnolence de Karolinska – KSS (Åkerstedt et Gillberg, 1990), fournie

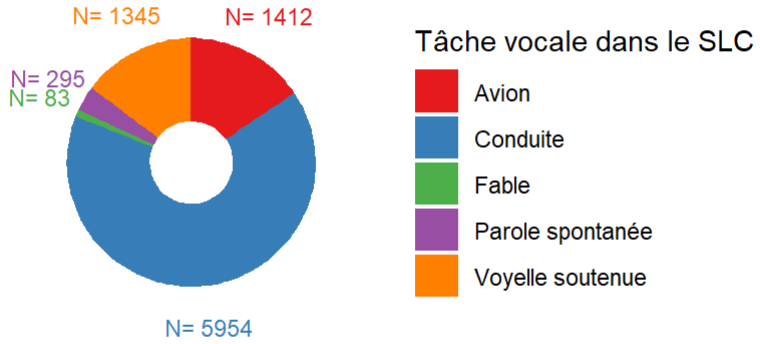

FIGURE 6.1 – Distribution des échantillons du SLC selon les différentes tâches vocales proposées.

en Annexe C. Cette échelle existe en deux versions : la version originale du questionnaire avec des descriptions textuelles uniquement sur les scores impairs, et une avec des étiquettes sur tous les échelons. Une étude rigoureuse a montré que les deux questionnaires sont équivalents et peuvent être utilisés de manière interchangeable (Miley *et coll.*, 2016).

Contrairement à l'échelle présentée dans le chapitre 4 qui est un autoquestionnaire rempli uniquement par le patient, l'annotation dans le SLC est mixte : la valeur retenue est la moyenne d'une KSS remplie par le patient lui-même, et de deux KSS remplies par des annotateurs externes entraînés à reconnaître la somnolence (Schuller *et coll.*, 2011). Cette mesure, qui n'a pas de validité médicale à notre connaissance, serait une mesure mixte de la somnolence subjective du sujet et des manifestations comportementales de son niveau de somnolence objectif et subjectif.

L'approche la plus commune trouvée dans la littérature pour estimer la somnolence dans la voix est de simplifier le problème en classification binaire. Pour cela, un seuil est défini pour distinguer les échantillons ayant été enregistrés par des locuteurs somnolents – S – et les locuteurs non somnolents – NS. En se basant sur une expérience détaillée dans (Krajewski *et coll.*, 2009), le seuil de référence proposé durant la compétition IS11 est fixé à 7.5, seuil en dessous duquel aucun épisode de microsommeil n'a été observé durant des tâches vocales surveillées par EEG. De plus, une étude approfondie de la KSS (Kaida *et coll.*, 2006) a conclu qu'un score supérieur à 7 est un bon indicateur d'un état d'éveil et de vigilance altéré, confortant le choix précédent.

6.1.3 Métadonnées

Dans le SLC, les seules métadonnées disponibles en dehors du label sont l'identifiant du locuteur et son sexe. Aucune autre métadonnée concernant les locuteurs n'est fournie. Ces données sont présentées dans le tableau 6.1.

6.2 SLEEP

Enregistré dans les mêmes laboratoires que le SLC, le corpus SLEEP (aussi appelé *Düsseldorf Sleepy Language Corpus*) a été rendu public en 2019 durant la compétition sur l'estimation continue de la somnolence proposée lors de la conférence Interspeech 2019 – IS19 (Schuller *et coll.*, 2019). Le principal avantage de ce corpus réside dans le nombre de locuteurs qui ont été enregistrés (915), produisant un total de 16464 échantillons.

SLEEPY LANGUAGE CORPUS – SLC				
LOCUTEURS				
Hommes	43			
Femmes	56			
Échantillons/locuteur (é-t)	91.8 (146.7)			
ÉCHANTILLONS				
	S	**NS**	**TOTAL**	**sig.**
Durée totale	6h 6min 9s	15h 10min 39s	21h 16min 48s	
Durée moyenne d'un échantillon (é-t)	7.0s (11.3s)	9.2s (17.0s)	8.15s (15.3s)	MW : ****
Échantillons	3137	5952	9089	
Hommes	716	1974	2690	χ^2 : ****
Femmes	2421	3976	6397	
KSS moy. (é-t)	8.33 (0.57)	4.35 (1.78)	5.72 (2.41)	

TABLEAU 6.1 – Statistiques du *Sleepy Language Corpus*. *S* : Somnolent (KSS $\geq$ 7.5). *NS* : Non-Somnolent (KSS $<$ 7.5). *sig.* : significativité des tests statistiques. *MW* : test de Mann-Whitney. **** : $p < 0.0001$.

6.2.1 Population et tâche vocale

Comme aucune information n'est disponible sur les sujets enregistrés, nous supposons, comme dans le SLC, qu'ils ont été choisis dans la population générale et que leurs éventuels problèmes de sommeil ont été mesurés à l'aide d'un questionnaire PSQI.

Les sujets ont été enregistrés durant des sessions qui durent entre 15 minutes et 1 heure. Les tâches vocales comprennent différents passages de lecture et des tâches de parole spontanée. Ces dernières ont été enregistrées en demandant aux locuteurs de commenter un évènement de leur vie (par exemple leur dernier week-end ou le meilleur cadeau qu'ils ont reçu) ou de décrire une image. Malheureusement, aucune information plus précise n'est donnée dans l'article introduisant la compétition, et cette information n'est pas présente dans la base de données.

De même que pour le SLC, le corpus SLEEP est divisé en sous-corpus d'entraînement, de développement et de test. L'annotation du sous-corpus de test n'est pas disponible publiquement : nous présentons dans cette partie uniquement les caractéristiques des sous-corpus d'entraînement et de développement.

6.2.2 Mesure de la somnolence

Dans le corpus SLEEP, la mesure de la somnolence est également la moyenne de trois KSS. Contrairement au SLC, la version utilisée est celle allant de 1 à 9, et les valeurs moyennes sont tronquées à l'entier le plus proche.

6.2.3 Métadonnées

Aucune métadonnée n'est disponible sur ce corpus : les seules informations incluses dans la base de données sont les échantillons audio et leur annotation avec la KSS.

Seul l'âge est précisé dans l'article introduisant la base de données (Schuller *et coll.*, 2019) : entre 12 et 84 ans, avec un âge moyen de 27.6 ± 11.0 ans.

Les données disponibles sur ce corpus sont présentées dans le tableau 6.2.

CORPUS SLEEP				
LOCUTEURS				
Hommes	551			
Femmes	364			
ÉCHANTILLONS				
	S	**NS**	**TOTAL**	**sig.**
Durée totale	3h 17min 1s	8h 24min 15s	11h 41min 17s	
Durée moyenne d'un échantillon (é-t)	3.84 (0.66)	3.88 (0.64)	3.87 (0.64)	MW : *
Nombre d'échantillons (ent. + dév.)	3081	7811	10892	
KSS (é-t)	7.61 (0.66)	4.07 (1.4)	5.07 (2.01)	

TABLEAU 6.2 – Statistiques des sous-corpus d'entraînement et de développement du corpus SLEEP. *S* : Somnolent (KSS $\geq$ 7). *NS* : Non-Somnolent (KSS $<$ 7). *sig.* : significativité des tests statistiques. *MW* : test de Mann-Whitney. * : $p < 0.05$.

6.3 Test de Maintien de l'Éveil – base TME

À notre connaissance, la base TME a été la première tentative de constituer une base de données de grande taille d'enregistrements vocaux de sujets pathologiques annotés par un test médical validé et reposant sur des mesures polysomnographiques (PSG).

En raison de nombreux biais et données incomplètes, ce corpus et les résultats préliminaires obtenus avec n'ont pas été publiés. Cependant, il a ouvert la voie à d'autres corpus, comme la base TILE présenté dans le prochain paragraphe. De plus, il possède des caractéristiques intéressantes qui permettent de nourrir la discussion menée dans le chapitre 9.

6.3.1 Population et tâche vocale

Ce corpus a été enregistré à la clinique du Sommeil du Centre Hospitalier Universitaire de Bordeaux en incluant 75 patients venus passer un Test de Maintien de l'Éveil – TME (*Maintenance of Wakefulness Test* – MWT en anglais). Ces patients se plaignent de somnolence diurne excessive et sont suspectés d'avoir des baisses de vigilance durant la journée.

Avant chacune des quatre itérations du TME, les patients lisent à voix haute un texte qui est soit le résumé d'un article de vulgarisation scientifique, soit une fable. Les enregistrements sont réalisés grâce au microphone intégré à une webcam, à une distance approximative de 30cm de la bouche des patients.

Ces textes sont inclus dans l'Annexe D et une brève description de ceux-ci (nombre de mots et durée moyenne des enregistrements correspondants) est proposée dans le tableau 6.3.

Session	Texte	Longueur	Durée moy. $\pm$ é-t
1	Des rats tenant conseil	196 mots	75.5 $\pm$ 15.0s
2	Pourquoi ne faut-il surtout pas boire de l'eau de mer pour s'hydrater ?	282 mots	112.0 $\pm$ 18.4s
3	Des lunettes providentielles	163 mots	60.1 $\pm$ 12.0s
4	La science vous donne une excellente raison de manger du chocolat !	278 mots	123.5 $\pm$ 26.2s

TABLEAU 6.3 – Caractéristiques des textes utilisés dans la base TME.

En raison des données manquantes pour de nombreux patients, les statistiques à l'échelle des locuteurs ne sont calculées que pour les 57 locuteurs pour lesquels les quatre valeurs de latence d'endormissement au TME sont collectées. Les statistiques à l'échelle des échantillons sont calculées sur l'ensemble des 75 locuteurs inclus dans le corpus.

6.3.2 Mesure de la somnolence – TME

Description du TME

Le Test de Maintien de l'Éveil – TME – est un des tests de référence pour mesurer de manière objective l'altération de la vigilance diurne (Mitler *et coll.*, 1982; Arand *et coll.*, 2005).

Durant ce test médical, les patients doivent résister au sommeil lorsqu'ils sont installés dans un fauteuil, avec une lumière tamisée, durant 40 minutes. Le test est divisé en quatre périodes séparées chacune de deux heures (10h, 12h, 14h et 16h). Un schéma explicatif du TME est proposé dans la figure 6.2.

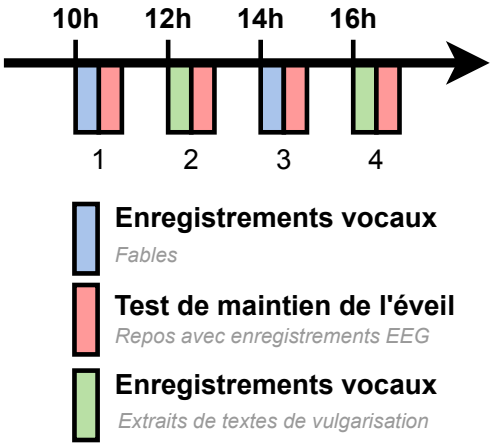

FIGURE 6.2 – Inclusion de la tâche de lecture dans le Test de Maintien de l'Éveil.

Si les patients s'endorment lors du test, ils sont immédiatement réveillés. Entre les différentes itérations du test, les patients doivent rester éveillés et sont libres de faire n'importe quelle activité exceptée du sport. Ils doivent arrêter de fumer 30 minutes avant le début de chaque itération du test et la consommation de café, thé ou toute autre boisson stimulante est interdite.

Mesure médicale

Durant le test, des signaux EEG, EMG et EOG sont enregistrés et annotés par des spécialistes, permettant de mesurer le temps d'endormissement du patient. La durée entre le moment où les patients sont mis au lit et la lumière éteinte, et le moment où ils s'endorment (au moins une page de n'importe quelle phase de sommeil) est appelée *latence d'endormissement au TME*.

La mesure médicale de référence mentionnée dans la section 4 est la valeur moyenne des latences d'endormissement sur les quatre siestes. Chaque session a une durée maximale de 40 minutes : les sessions durant lesquelles les patients ne se sont pas endormis sont annotées avec une valeur de 40 minutes de latence d'endormissement.

Pour la simplification de l'annotation en deux classes – S et NS – le seuil est habituellement fixé à 19 minutes sur la latence moyenne d'endormissement (Doghramji *et coll.*, 1997; Sagaspe *et coll.*, 2007; Arand *et coll.*, 2005).

Annotation dans la base TME

Les échantillons de la base TME sont annotés de deux manières différentes : avec la latence moyenne d'endormissement pour l'estimation de trait des locuteurs ; et avec chaque latence individuelle pour la détection d'état au court terme. Les latences d'endormissement prises de manière individuelle ne sont pas des mesures validées médicalement, mais peuvent être vues comme des marqueurs objectifs des variations à court terme du niveau de vigilance du sujet au cours des sessions. Les échantillons sont classés entre les classes S et NS avec le même seuil de 19 minutes que pour les latences moyennes d'endormissement, indépendamment du label du locuteur qui a été enregistré.

6.3.3 Métadonnées

Les métadonnées de la base TME sont les ID des locuteurs, leur âge, sexe, indice de masse corporelle, diagnostic pour l'apnée du sommeil, et score à l'échelle de somnolence d'Epworth [ESS,(Johns, 1991)]. L'échelle d'activation de l'hôpital de Toronto [THAT, (Shahid *et coll.*, 2016)] et l'échelle de somnolence en visages [CFS, (Maldonado *et coll.*, 2004)] ont aussi été remplies par quelques patients, mais pas suffisamment pour permettre une analyse pertinente.

Les données disponibles sur ce corpus sont présentées dans le tableau 6.4.

BASE TME				
LOCUTEURS				
	S	**NS**	**TOTAL**	**sig.**
Locuteurs	11	46	57	
Âge (é-t) années	45.9 (15.6)	46.2 (15.9)	46.14 (15.72)	MW : n.s.
IMC (é-t) kg/m^2	30.2 (4.4)	18.5 (13.0)	20.8 (12.4)	MW : **
SAOS	9	25	34	χ^2 : n.s.
Sans SAOS	2	20	22	
Hommes	10	29	39	χ^2 : n.s.
Femmes	1	17	18	
TME (é-t) minutes	10.36 (4.9)	34.1 (7.21)	29.53 (11.64)	
ESS (é-t)	13.44 (4.53)	10.93 (5.24)	11.38 (5.17)	MW : n.s.
ÉCHANTILLONS				
	S	**NS**	**TOTAL**	**sig.**
Durée totale	2h 04min 14s	5h 08min 36s	7h 12min 51s	
Durée moy. (é-t)	1min 33s (31.9s)	1min 30s (31.3s)	1min 32s (31.7s)	MW : n.s.
Échantillons	83	199	282	
Hommes	63	126	189	χ^2 : n.s.
Femmes	20	73	93	
TME (é-t) minutes	6.37 (4.12)	38.52 (4.37)	29.06 (15.3)	

TABLEAU 6.4 – Statistiques de la base TME. *S* : Somnolent (TME moy. $\leq$ 19 minutes et TME $\leq$ 19 minutes pour le label resp. des locuteurs et des échantillons). *NS* : Non-Somnolent. *sig.* : significativité des tests statistiques. *SAOS* : Syndrome d'Apnée Obstructive du Sommeil. *MW* : test de Mann-Whitney. *n.s.* : non significatif, ** : $p < 0.01$.

Chapitre 7

Base TILE

Sommaire

La base TILE a été élaborée dans la continuité de la base TME. Également enregistré au service universitaire de médecine du Sommeil du Centre Hospitalier Universitaire de Bordeaux, ce corpus poursuit l'objectif de la base TME de trouver des biomarqueurs vocaux de la somnolence objective chez une population pathologique.

Ce chapitre présente les caractéristiques du corpus TILE-106 (cf Section 7.5).

7.1 Population et tâche vocale

122 patients de la clinique du Sommeil ont été enregistrés lors de leur passage d'un Test Itératif de Latence d'Endormissement – TILE – pour diagnostic ou suivi d'une maladie du sommeil. Ce test consiste à demander aux patients de faire une courte sieste cinq fois dans la journée. Avant chaque sieste, les sujets remplissent une échelle de somnolence de Karolinska et sont enregistrés en train de lire un texte d'environ 200 mots.

Les patients sont généralement enregistrés assis à leur bureau ou sur leur lit. Aucun enregistrement n'a été effectué avec des patients allongés. Les textes lus sont extraits du Petit Prince d'Antoine de Saint-Exupéry. Ces textes sont proposés en Annexe D et leurs caractéristiques (nombre de mots et durée moyenne des enregistrements correspondants) sont présentées dans le tableau 7.1.

Session	Texte	Longueur	Durée moy. $\pm$ é-t
1	Texte 1	231 mots	77.9 $\pm$ 11.6s
2	Texte 2	235 mots	79.3 $\pm$ 12.2s
3	Texte 3	228 mots	72.7 $\pm$ 10.8s
4	Texte 4	221 mots	77.2 $\pm$ 12.2s
5	Texte 5	257 mots	80.6 $\pm$ 12.4s

TABLEAU 7.1 – Caractéristiques des textes utilisés dans la base TILE et durée de lecture des patients du sous-corpus TILE-106.

Un sixième texte (Texte 0) est inclus à l'Annexe D : il s'agit du texte lu la veille du TILE, après l'arrivée des patients au service universitaire de médecine du Sommeil, pour leur présenter l'étude et faire un premier enregistrement qui sert à la fois à évaluer leur niveau de lecture et les habituer à la tâche de lecture à voix haute et au matériel. Les enregistrements effectués lors de cette « Session 0 » ne sont inclus dans aucune des analyses proposées dans ce document.

7.2 Critères d'inclusion et d'exclusion

Une centaine de sujets ont dans un premier temps été inclus, avant la collaboration avec Mathilde Rieant et Gabrielle Chapoutier (CFUOB – Université de Bordeaux). En plus d'un nouveau type de marqueur pouvant être pertinent au regard de la somnolence (les erreurs de lecture, cf. chapitre 13), ce travail avec des élèves en master d'orthophonie a permis de mettre au jour un biais jusqu'alors inexploré dans les bases de données sur la détection des pathologies neuropsychiatriques dans la parole : le niveau de lecture et d'élocution du sujet.

Dans un premier temps, nous avons écouté et évalué le niveau de lecture des patients déjà inclus dans la base TILE. Quatorze patients ont ainsi été exclus du corpus a posteriori en raison d'un comportement anormal de lecture : les différents critères d'exclusion sont présentés dans le tableau 7.2.

Ensuite, nous avons conçu un ensemble de critères d'exclusion portant sur le niveau de lecture des patients à vérifier lors de la visite de test, la veille du test médical. Ces critères sont les suivants :

— Langue maternelle du sujet autre que le français ;
— Antécédents d'Accident Vasculaire Cérébral ou d'Accident Ischémique Transitoire, le comportement de lecture pouvant être dû à des séquelles de ceux-ci (alexie, troubles visuoattentionnels) ;
— Troubles de la fluence (bredouillement, bégaiement), de la parole, et de l'articulation ;
— Maladies neuromusculaires ;
— Trouble déficitaire de l'attention avec ou sans hyperactivité (TDAH) ;
— Trouble du langage écrit.

Une grille d'évaluation portant sur la lecture durant cette session est proposée dans le tableau 7.2.

Deux patients supplémentaires n'ont pas été enregistrés sur les 5 itérations du TILE, et ont donc également été exclus du corpus.

7.3 Mesure de la somnolence – TILE

7.3.1 Description du TILE

Le Test Itératif de Latence d'Endormissement est la mesure de référence de la somnolence diurne excessive (Carskadon et Dement, 1977; Arand *et coll.*, 2005). Il diffère du TME par la dimension de la somnolence qu'il mesure : le TILE est une mesure objective de la propension à l'endormissement, dans un contexte de SDE. Concrètement, le test consiste à demander aux sujets de faire cinq siestes durant la journée, à 9h, 11h, 13h, 15h et 17h. Un schéma explicatif de la procédure est proposé figure 7.1.

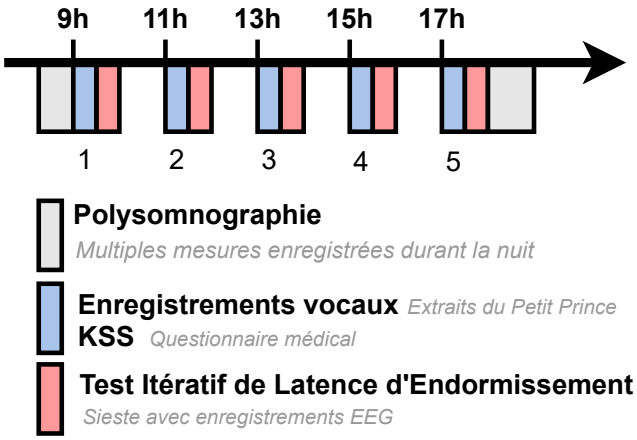

FIGURE 7.1 – Inclusion de la tâche de lecture dans le Test Itératif de Latence d'Endormissement.

Les sujets ont 20 minutes pour s'endormir, et chaque sieste a une durée maximale de 35 minutes : si le patient s'endort, sa sieste est éventuellement allongée de façon à ce qu'il dorme 15 minutes au total. La procédure est la même que pour le TME, et la durée entre le début du test et l'endormissement est appelée *latence d'endormissement au TILE*. Contrairement au TME, où l'on demande aux sujets de lutter contre le sommeil et de rester éveiller, la consigne pour le TILE est de « se relaxer et de se laisser aller au sommeil ». Concernant les périodes entre

Raison d'exclusion	N
Comportement de lecture	
Lecture trop lente	1
Suspicion de dyslexie	3
Saute une ligne	2
Bredouilleur	1
Pathologie des cordes vocales	
Dysphonie spasmodique	1
Autres pathologies	
Rupture d'anévrisme	1
TDAH	1
Cocktail médicamenteux	2
AIT	1
Myotonie de Steinert	1
Total	14

TROUBLES DE LA LECTURE
Lenteur de lecture + impression de déchiffrage
Lecture rapide, mais beaucoup d'erreurs d'adressage (paralexies)
Erreurs souvent non corrigées
Beaucoup de paralexies

TROUBLE DE LA FLUENCE (bégaiement, bredouillement...)
Blocage/allongements/répétitions fréquents
Débit très rapide
Nombreux télescopages

TROUBLE DE L'ATTENTION
Saut de ligne/répétition d'une ligne
Nombreux oublis
Lenteur et/ou nombre élevé d'erreurs

TROUBLE VISUEL
Saut de ligne/répétition d'une ligne
Nombreux oublis
Paralexies nombreuses

TROUBLE DE LA PAROLE ET/OU D'ARTICULATION
Difficultés de mise en séquence des sons dans la chaîne de locution
Phonétisme incomplet (ex. réalisation impossible du /ch/) ou atypique (ex. « zozotement »)

Tableau 7.2 – Patients exclus a posteriori de la base TILE en raison de pathologies interférant avec leurs capacités de lecture (gauche) et grille d'évaluation du niveau de lecture des patients de la base TILE lors de la visite la veille du test médical (droite).

les siestes, les mêmes instructions que pour le TME s'appliquent. Pour la procédure médicale exacte du TILE, nous redirigeons le lecteur vers (Littner *et coll.*, 2005).

7.3.2 Mesure médicale

En se basant sur l'interprétation des signaux EEG, EMG et EOG, les spécialistes entraînés peuvent déterminer le moment où le patient s'endort. La latence d'endormissement au TILE a une valeur maximale de 20 minutes : les patients qui ne se sont pas endormis sont annotés avec cette valeur seuil. La valeur de référence mentionnée dans le chapitre 4 est la valeur moyenne des cinq latences d'endormissement au cours du test.

Le TILE est utilisé dans le diagnostic de nombreuses pathologies, dont la narcolepsie dont il est le test de diagnostic de référence. Pour cette dernière pathologie, le seuil de diagnostic a été fixé à 8 minutes (Aldrich *et coll.*, 1997) : en dessous de cette valeur, la propension à l'endormissement est considérée comme pathologique. L'*American Association of Sleep Medecine*

a élargi l'utilisation de ce seuil dans un rapport de 2005 (Arand *et coll.*, 2005). Nous utilisons donc ce seuil pour discriminer les patients Somnolents – c'est à dire qui ont une propension à l'endormissement considérée comme pathologique – et les patients Non Somnolents.

7.3.3 Annotation dans la base TILE

De même que pour la base TME, nous annotons les échantillons à la fois avec la latence d'endormissement de chaque sieste et la latence moyenne, indépendamment l'une de l'autre. Le seuil utilisé pour les deux annotations précédentes est celui de 8 minutes mentionné dans le paragraphe précédent. Les latences d'endormissement à chaque sieste ne sont pas des mesures validées médicalement, mais peuvent être des marqueurs objectifs de la propension à l'endormissement à court terme dans des conditions favorisant le sommeil.

7.4 Métadonnées

Par rapport aux trois précédents corpus, celui-ci contient de très nombreuses métadonnées concernant l'état de santé des patients inclus, comme des données physiques (taille, poids, IMC, tour de cou), mais aussi des questionnaires de fatigue, d'insomnie, de qualité de vie, etc. L'intégralité des informations collectées sur les patients dans la base TILE est présentée dans le tableau 7.3, dont les acronymes et les amplitudes de valeurs possibles sont présentés dans le tableau 7.4.

Cette base de données est la première collectant également les pathologies des patients inclus, qui sont représentées dans la figure 7.2.

7.5 Différentes versions du corpus TILE

Au cours de ce manuscrit, différentes versions de la base TILE seront exploitées. Afin d'être les plus explicites possibles sur ce point qui peut être source de confusion, nous dénotons dans la suite les différentes versions par le nombre de locuteurs qu'elles incluent.

7.5.1 base TILE-122

Cette version de la base TILE correspond à tous les enregistrements qui ont été effectués pour le corpus. Elle contient donc les enregistrements des 14 patients qui ont des troubles de la lecture et les enregistrements des deux patients qui n'ont pas été enregistrés sur les cinq itérations du TILE.

7.5.2 base TILE-106

Cette version du corpus correspond à la version précédente à laquelle nous avons exclu les patients atteints de pathologies pouvant affecter la lecture de textes et ceux qui n'ont pas été enregistrés sur les cinq siestes. C'est cette version de la base TILE qui sera utilisée dans le chapitre suivant, et qui a été utilisée pour les résultats présentés dans les chapitres 11 et 12.

7.5.3 base TILE-93

Lorsque les pathologies du sommeil affectant les patients ont été incluses a posteriori dans le corpus. Nous avons alors élaboré un nouveau sous-corpus, la base TILE-93, qui est

un sous-corpus de la base TILE-106 pour lequel nous avons exclu les patients non diagnostiqués ($n = 4$) et les patients atteints de narcolepsie ($n = 12$), ces derniers ayant des latences moyennes d'endormissement au TILE significativement plus faible que les autres pathologies (cf. figure 8.12 dans le chapitre suivant). C'est cette version du corpus qui est utilisée à partir du chapitre 14. L'intégralité des informations de cette version du corpus est proposée dans l'annexe E.

Caractéristique	TILE moy. $\leq$ 8.0 min	TILE moy. $>$ 8.0 min	Tous
Caractéristiques physiques et sociodémographiques			
Sexe	F : 13.0 M : 15.0	F : 50.0 M : 28.0	F : 63.0 M : 43.0
Âge	33.5 (15.6) ***	36.7 (13.4) ***	35.9 (14.1)
Taille (m)	1.7 (0.1) *	1.7 (0.1) *	1.7 (0.1)
Poids (kg)	70.4 (15.5)	69.9 (17.0)	70.0 (16.6)
IMC (kg/m^2)	24.1 (4.6)	24.3 (5.5)	24.2 (5.3)
Tour de cou (cm)	38.7 (3.7) **	37.6 (4.5) **	37.9 (4.3)
Niveau d'éducation	4.2 (2.0) ***	5.8 (2.6) ***	5.4 (2.5)
Somnolence à court terme			
TILE (min.)	4.8 (3.5) ***	13.5 (5.8) ***	11.2 (6.5)
KSS (1-9)	4.3 (1.8)	4.5 (1.9)	4.4 (1.9)
Visage (0-4)	1.6 (0.8)	1.6 (0.9)	1.6 (0.9)
Mesures de somnolence excessive			
TILE moy. (min.)	4.8 (2.0) ***	13.5 (3.2) ***	11.2 (4.8)
ESS (0-24)	15.9 (5.6) ***	14.2 (4.5) ***	14.6 (4.9)
BSI (0-6)	2.5 (1.2)	2.3 (1.0)	2.3 (1.0)
Dimensions liées à la somnolence			
THAT (0-50)	24.2 (8.2) *	23.0 (7.1) *	23.3 (7.5)
Hobson (0-16)	4.8 (2.9) **	3.9 (2.2) **	4.1 (2.5)
ISI (0-28)	14.6 (6.3) *	15.5 (5.2) *	15.3 (5.5)
ASRS (0-24)	12.8 (5.1) *	12.3 (5.1) *	12.4 (5.1)
FOSQ-10 (10-40)	21.4 (7.0)	21.0 (6.5)	21.1 (6.7)

FSS (9-63)	45.8 (13.1) *	49.4 (9.8) *	48.4 (10.9)
Fatigue	0 : 4.0 1 : 24.0	0 : 6.0 1 : 72.0	0 : 10.0 1 : 96.0
Durée de sommeil estimé la veille (h)	7.2 (2.1)	7.2 (1.8)	7.2 (1.9)
Ronflements	0 : 22.0 1 : 6.0	0 : 60.0 1 : 18.0	0 : 82.0 1 : 24.0
Apnées observées	0 : 23.0 1 : 5.0	0 : 61.0 1 : 17.0	0 : 84.0 1 : 22.0
Hypertension	0 : 21.0 1 : 7.0 *	0 : 72.0 1 : 6.0 *	0 : 93.0 1 : 13.0
Pathologie	H Non spécifiée : 2.0 H.NEURO : 3.0 H.SAOS : 4.0 H.TDAH : 3.0 HI : 9.0 NT1 : 1.0 NT1/SAOS : 1.0 NT2 : 5.0	H Non spécifiée : 18.0 H.NEURO : 3.0 H.SAOS : 14.0 H.TDAH : 16.0 H.Tb Psy : 3.0 HI : 17.0 HI : 1.0 INTROUVABLE : 4.0 NT1 : 1.0 NT2 : 1.0	H Non spécifiée : 20.0 H.NEURO : 6.0 H.SAOS : 18.0 H.TDAH : 19.0 H.Tb Psy : 3.0 HI : 26.0 HI : 1.0 INTROUVABLE : 4.0 NT1 : 2.0 NT1/SAOS : 1.0 NT2 : 6.0
Anxiété et Dépression			
HAD-D (0-21)	5.6 (4.0) ***	6.6 (3.8) ***	6.3 (3.9)
HAD-A (0-21)	8.0 (3.8)	8.5 (4.1)	8.4 (4.0)
Addictions			
CAGE (0-4)	0.4 (0.9) *	0.5 (0.9) *	0.5 (0.9)
CDS5 (5-25)	6.9 (4.5)	6.9 (4.3)	6.9 (4.4)
Nb cigarettes par sem.	1.9 (4.9)	1.9 (5.1)	1.9 (5.0)
Nb alcool par sem.	0.2 (0.4) ***	0.4 (0.8) ***	0.3 (0.7)

TABLEAU 7.3 – Caractéristiques des patients du corpus TILE-106. Orange : différence significative (test de Mann-Whitney). Vert : différence significative (test du χ^2).

Mesure	Référence	Description
Caractéristiques physiques et sociodémographiques		
Niveau d'éducation	-	Nombre d'années d'études après le brevet des collèges
Somnolence à court terme		
Latence d'endormissement au TILE	(Littner *et coll.*, 2005)	Temps (en min.) entre le début du test et l'endormissement du patient (0-20 min)
KSS	(Åkerstedt et Gillberg, 1990)	1 item sur la somnolence au cours des 10 dernières min. (1-9)
Échelle des visages	(Maldonado *et coll.*, 2004)	5 dessins de visages mesurant la somnolence (0-4)
Mesures de somnolence excessive		
Latence moyenne d'endormissement au TILE	(Littner *et coll.*, 2005)	Moyenne des cinq latences d'endormissement (0-20 min.)
Epworth Sleepiness Scale (ESS)	(Johns, 1991)	8 items à propos de la somnolence diurne (0-24)
Barcelona Sleepiness Index (BSI)	(Guaita *et coll.*, 2015)	2 items à propos de la somnolence (0-6)
Dimensions liées à la somnolence		
Toronto Hospital Alertness Test (THAT)	(Shahid *et coll.*, 2016)	10 items pour mesurer le niveau d'alerte (0-50)
Hobson	(Hobson *et coll.*, 2002)	4 items à propos de la somnolence diurne excessive (0-16)
Insomnia Severity Index (ISI)	(Bastien *et coll.*, 2001)	7 items à propos de l'insomnie (0-28)
Partie A de l'ADHD Self-Report Scale (ASRS)	(Schweitzer *et coll.*, 2001)	6 items à propos du TDAH (0-24)
Functional Outcomes of Sleep Questionnaire (FOSQ-10)	(Weaver *et coll.*, 1997)	10 items sur l'impact de la somnolence diurne excessive sur le fonctionnement quotidien (10-40)
Fatigue Severity Scale (FSS)	(Krupp *et coll.*, 1989)	9 items sur la fatigue (9-63)
Durée de sommeil estimé la veille	-	Durée de sommeil estimé par le patient lui-même lors de la nuit précédant le TILE (h)
Anxiété et dépression		
Hospital Anxiety and Depression scale	(Zigmond et Snaith, 1983)	7 items sur la dépression, 7 items sur l'anxiété (0-21 chacun)
Addictions		
Cut-Down, Annoyed, Guilty, Eye-opener questionnaire (CAGE)	(Ewing, 1984)	4 items sur la consommation d'alcool
Cigarette Dependence Scale – version courte (CDS-5)	(Courvoisier et Etter, 2008)	5 items à propos de la dépendance à la cigarette

Tableau 7.4 – Informations collectées dans la base TILE.

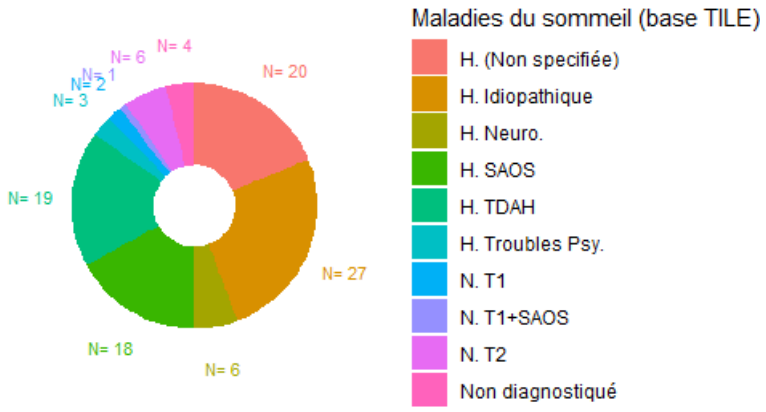

FIGURE 7.2 – Distribution des patients de la base TILE en fonction de leur diagnostic. H. : Hypersomnie, SAOS : Syndrome d'Apnée Obstructive d'Apnée du Sommeil, TDAH : Trouble Déficitaire de l'Attention avec ou sans Hyperactivité, Tb. Psy. : Troubles psychiatriques, N : Narcolepsie (Type 1 ou 2).

Chapitre 8

Comparaison des bases de données

Sommaire

À partir des informations présentées dans les deux précédents chapitres, nous proposons dans celui-ci une comparaison, élément par élément, des quatre corpus.

8.1 Tâches vocales

8.1.1 Durée des échantillons en fonction de la classe de somnolence

La durée des échantillons dans les quatre corpus est représentée dans la figure 8.1. À la fois dans le SLC et le corpus SLEEP, les échantillons étiquetés NS sont significativement plus longs que les ceux annotés comme S (test de Mann-Whitney, SLC : $p = 6.5 \times 10^{-10}$, SLEEP : $p = 0.03$). Au contraire, dans les bases TME et TILE, la taille des échantillons est la même dans les deux classes de somnolence.

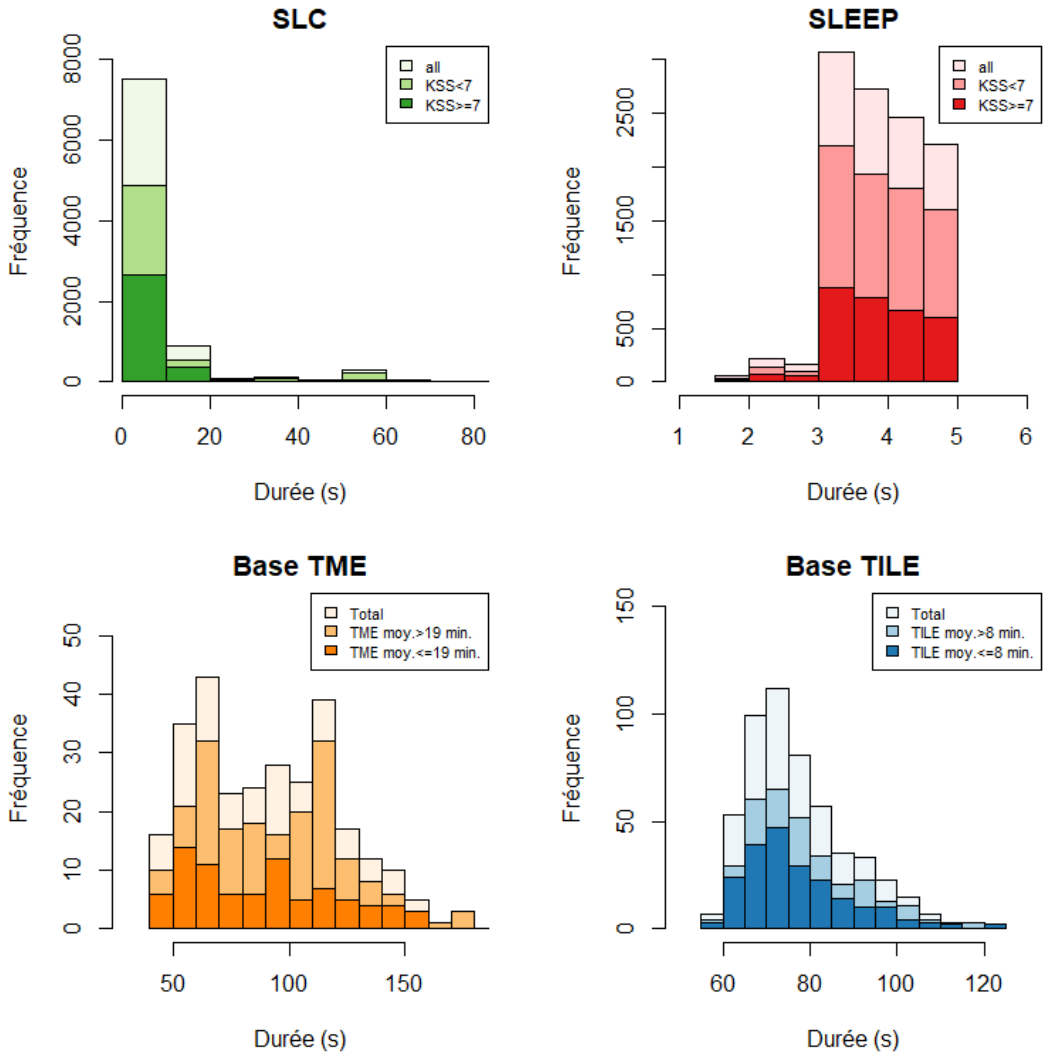

FIGURE 8.1 – Distribution de la durée des échantillons dans les bases SLC et SLEEP.

8.1.2 Durée en fonction de la tâche effectuée ou de l'itération

La taille des échantillons en fonction de la tâche dans le SLC est représentée dans la figure 8.2.

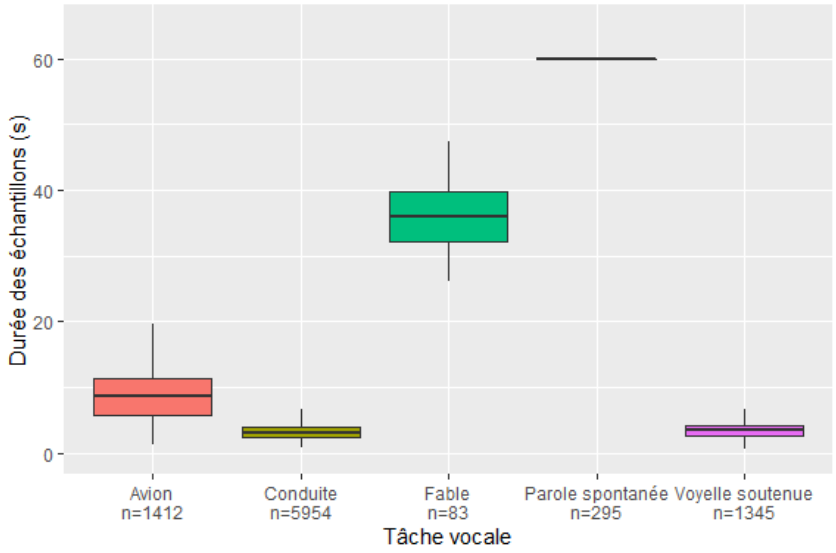

FIGURE 8.2 – Durée des échantillons du SLC en fonction de la tâche vocale.

Excepté les échantillons de parole spontanée, la majorité des échantillons ont une durée inférieure à 10s, notamment en raison de la très large représentation des échantillons contenant des voyelles soutenues ou des commandes automobiles, qui sont plus courts. Certaines tâches sont plus longues que les autres, comme par exemple les lectures en anglais qui ont une durée moyenne de 8.7 secondes ou la lecture de la fable qui fait plus du double, avec une longueur moyenne de 36.6 secondes.

Les longueurs des échantillons en fonction des textes lus dans la base TME et la base TILE sont représentées dans la figure 8.3.

Dans la base TILE, tous les enregistrements ont des tailles similaires (approximativement 75 secondes) excepté durant la troisième session, pour laquelle les textes sont plus courts. Notre étude portant sur une version préliminaire de la base TILE (Martin *et coll.*, 2020b) avait déjà fait cette observation, démontrant que cette observation n'est pas seulement due à la différence de taille entre les textes, mais aussi à une diminution de la KSS – c'est-à-dire une augmentation du niveau d'activation du locuteur, qui lit alors plus vite.

Au contraire, dans la base TME, toutes les sessions ont des tailles différentes. Cela s'explique principalement par le fait qu'il y a deux types de textes dans cette base de données : des fables durant les premières et troisièmes sessions, et des résumés d'articles de vulgarisation scientifique durant les deuxièmes et quatrièmes siestes.

8.1.3 Nombre d'échantillons par locuteur

Dans le SLC, en fonction de l'expérience à laquelle les sujets ont participé, ils sont enregistrés entre 3 et 909 fois. Dans la figure 8.4 (gauche), cinq locuteurs se distinguent de la distribution principale, en étant enregistrés un très grand nombre de fois. Le nombre moyen d'échantillons par locuteur a été recalculé sur SLC sans ces 5 cas particuliers, changeant la

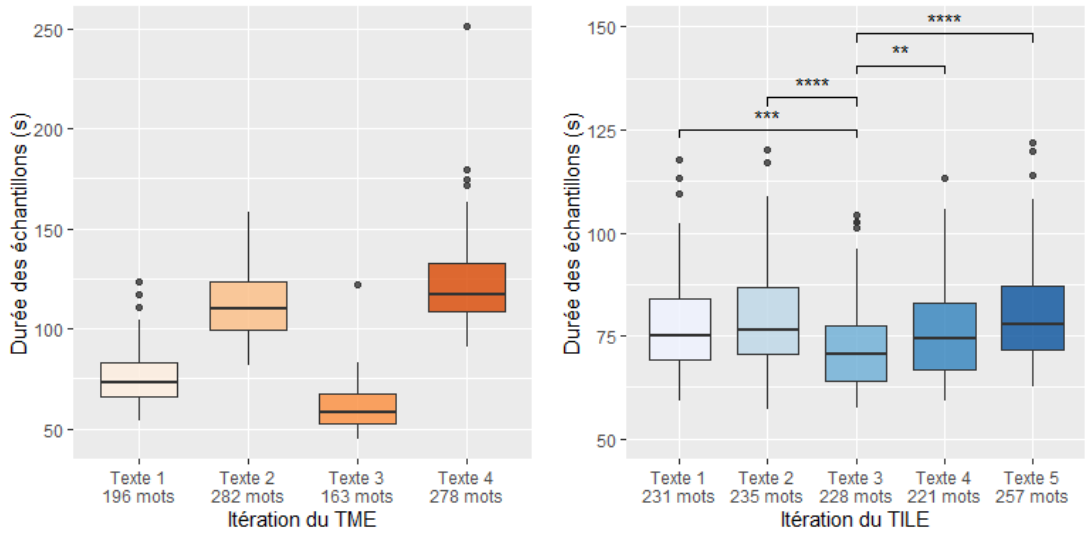

FIGURE 8.3 – Durée des échantillons dans les bases TME (gauche) et TILE (droite) en fonction du texte lu. La significativité des différences n'a pas été représentée dans la figure de gauche pour des raisons de lisibilité.

moyenne de 91 à 61 échantillons par sujet. L'histogramme sans ces 5 intrus est présenté dans la figure 8.4 (droite).

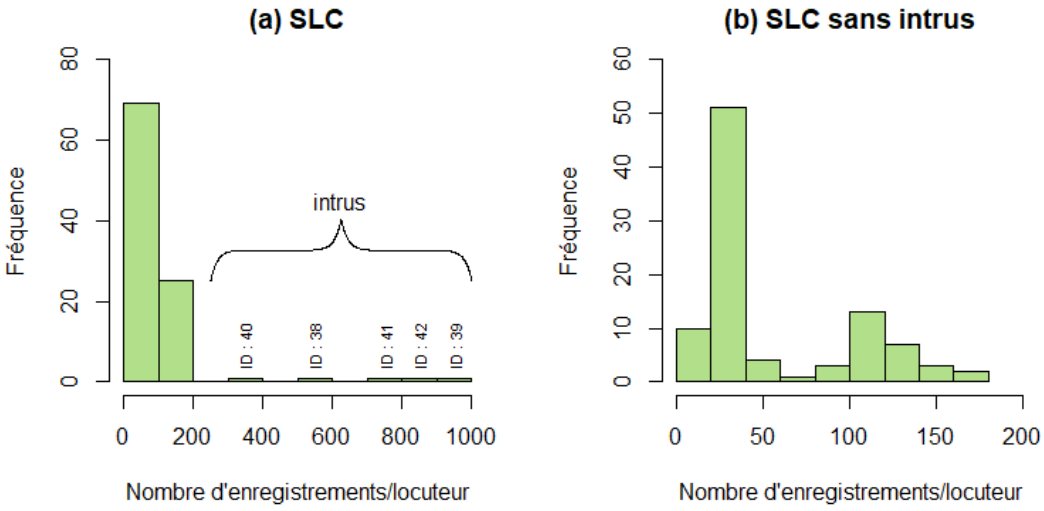

FIGURE 8.4 – Distribution du nombre d'enregistrements par locuteur dans le SLC avec (gauche) et sans (droite) les locuteurs ayant été enregistrés un très grand nombre de fois.

Dans le corpus SLEEP, les sessions d'enregistrements durent entre 15 minutes et une heure : le nombre d'échantillons par locuteur résultant de ces sessions est très certainement déséquilibré entre les locuteurs.

Enfin, dans la base TME et la base TILE, le nombre d'échantillons par locuteur est fixé par le nombre de sessions du test médical (4 pour le TME, 5 pour le TILE).

8.2 Annotations des échantillons

8.2.1 Latences d'endormissement au TME et au TILE

Sur les latences d'endormissement au TME et au TILE, une saturation est observée pour les patients qui ne s'endorment pas et à qui on assigne la valeur maximale du test (40 minutes pour le TME, 20 minutes pour le TILE). Nous employons ici le terme de *saturation* car ces valeurs ne déséquilibrent pas seulement les distributions globales des latences d'endormissement (cf. figure 8.5), mais elles représentent une grande partie des deux bases de données. En effet, les annotations correspondant à une saturation représentent 62% des échantillons de la base TME, et 24% des échantillons la base TILE.

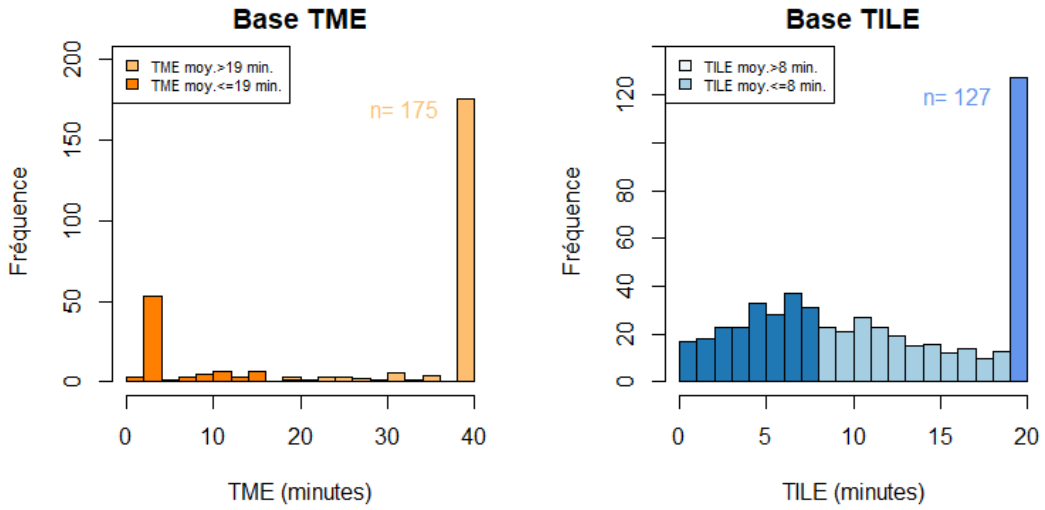

FIGURE 8.5 – Distribution des latences d'endormissement sur les bases TME (gauche) et TILE (droite). La barre la plus à droite représente le nombre d'échantillons annotés avec la valeur maximale de chaque test (40 minutes pour le TME, 20 minutes pour le TILE), représentant une saturation.

8.2.2 Latences d'endormissement moyennes au TME et au TILE

Les latences moyennes d'endormissement au TME et au TILE sont représentées dans la figure 8.6. Dans la base TME, la prédominance des latences de 40 minutes semble avoir un impact important sur la distribution des latences moyennes d'endormissement, qui présente la même saturation à 40 minutes.

Au contraire, dans la base TILE, très peu de patients ont une latence moyenne d'endormissement de 20 minutes. Les latences moyennes d'endormissement ont une distribution plus lisse, sans saturation.

8.2.3 Karolinska Sleepiness Scale

La KSS dans les corpus SLC, SLEEP et la base TILE sont représentés dans la figure 8.7. Ces mesures sont différentes : dans le SLC, il s'agit d'une moyenne de trois KSS allant de 1 à 10, l'un rempli par le patient, les deux autres par des annotateurs externes ; dans le corpus SLEEP, la procédure est la même avec la version allant de 1 à 9 et une moyenne tronquée ; et

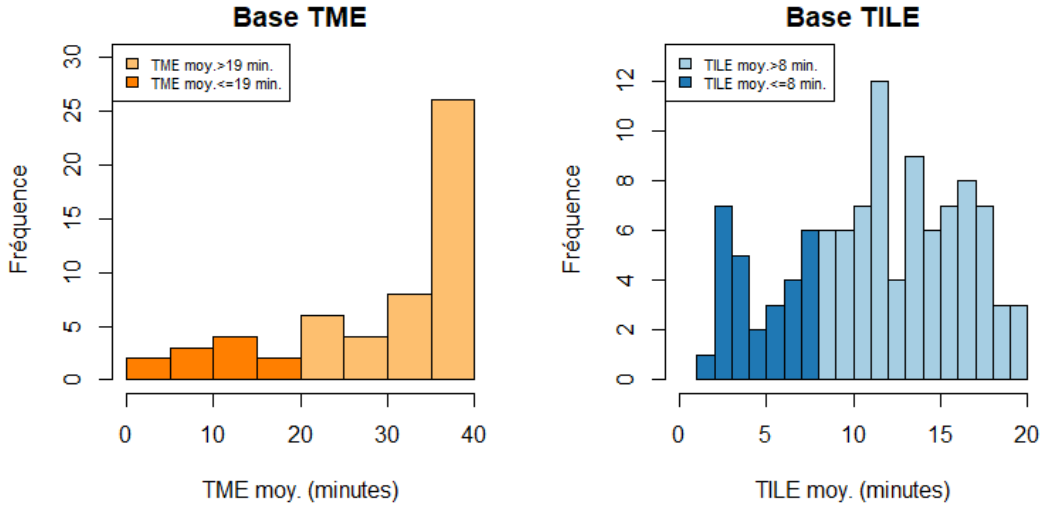

FIGURE 8.6 – Distribution des latences moyennes d'endormissement au TME (gauche) et TILE (droite) dans les bases de données respectives.

enfin, dans la base TILE, il s'agit de la même version que dans le corpus SLEEP, mais annoté uniquement par le locuteur.

Dans le SLC, les locuteurs qui ont été enregistrés un très grand nombre de fois ont une influence sur la distribution de la KSS dans le corpus. En effet, la différence de distributions des KSS du SLC avec et sans ces locuteurs, représentée dans la figure 8.8, est significative (MW, $p < 10^{-13}$). Cette différence induit des changements importants dans l'équilibre entre les classes S et NS : avec ces locuteurs, les échantillons S représentent 34% des 9089 échantillons, tandis que cette proportion chute à 27.5% des 5776 échantillons restants lorsqu'ils sont exclus du SLC.

La distribution de la KSS dans le corpus TILE diffère des deux autres corpus par une surreprésentation des scores impairs comparés aux scores pairs. L'item «3 – Éveillé» est celui ayant été le plus annoté, suivi par «5 – Ni éveillé, ni somnolent» et «7 – Somnolent, mais sans effort pour rester éveillé». Cette observation, courante en psychométrie, est expliquée par le fait que les patients annotent plus facilement leur état sur des niveaux qui possèdent une description textuelle que ceux intermédiaires, qui n'en possèdent pas.

Contrairement à ce qui a été observé dans le SLC en excluant les patients ayant été enregistrés de très nombreuses fois, exclure les échantillons qui correspondent à la saturation décrite précédemment n'a pas d'influence sur la distribution de la KSS dans la base TILE (cf figure 8.8). En effet, les deux distributions ne montrent pas de différences majeures (test de Mann-Whitney : $p = 0.91$).

8.2.4 Corrélation entre la KSS et la latence d'endormissement au TILE

La KSS et les valeurs de latence d'endormissement au TILE dans la base TILE ne corrèlent pas (ρ de Spearman entre la KSS et les latences d'endormissement au TILE, avec saturations : $\rho = -0.034$). Nous avons d'abord pensé que cette observation était due aux valeurs de saturation des latences d'endormissement, mais en excluant ces échantillons, la corrélation augmente de façon négligeable (ρ de Spearman entre la KSS et les latences d'endormissement au TILE, sans saturations : $\rho = -0.042$). Il semble ainsi y avoir une différence intrinsèque entre

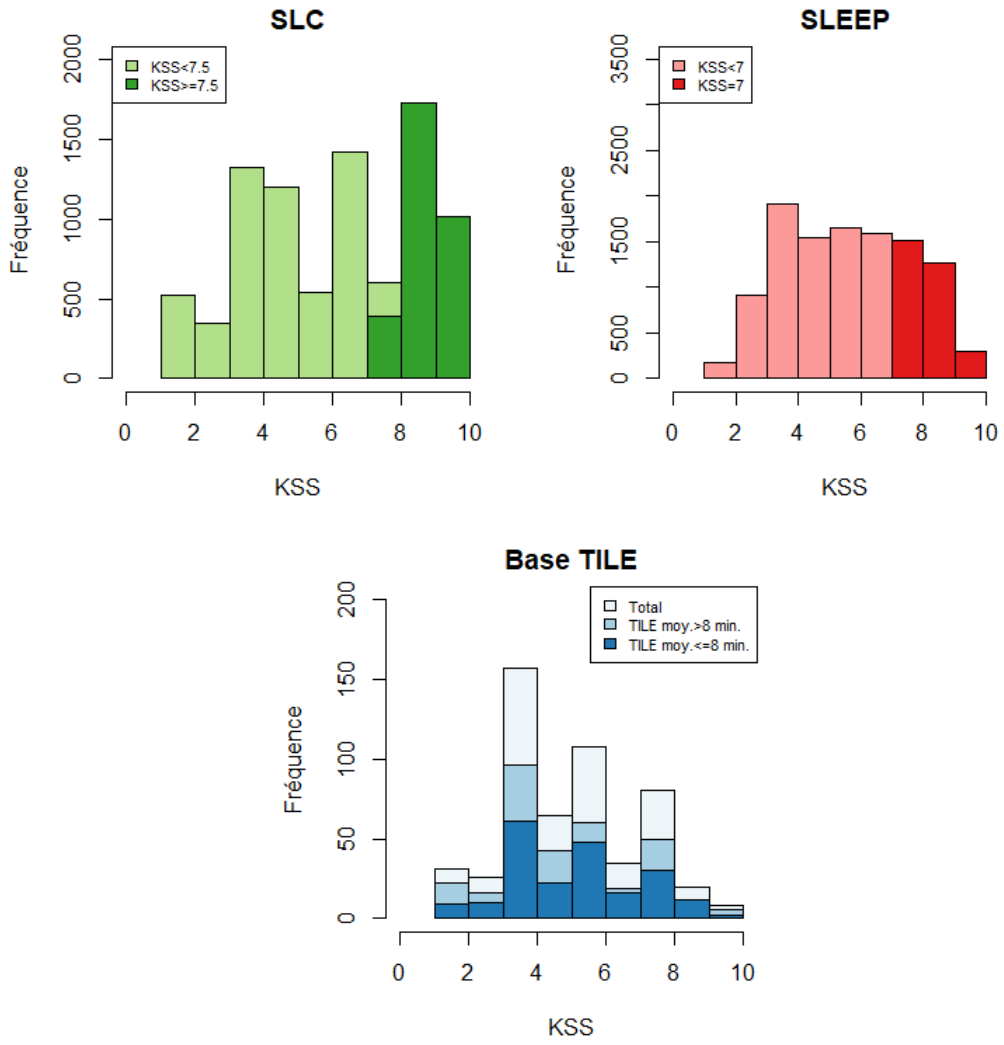

FIGURE 8.7 – Distribution de la KSS dans le SLC (haut gauche), le corpus SLEEP (haut droit) et la base TILE (bas).

ce que mesurent les latences d'endormissement au TILE et le construit mesuré par la KSS (cf. chapitre 5).

8.2.5 Epworth Sleepiness Scale

Dans la base TME, la distribution de l'ESS est quasiment significativement différente entre les deux classes de somnolence (test de Mann-Whitney, $p = 0.067$), tandis que dans la base TILE, cette différence est plus marquée (test de Mann-Whitney, $p = 0.04$). De plus, cette mesure est anti-corrélée de façon quasiment significative avec la latence d'endormissement moyenne au TILE (ρ de Spearman $\rho = -0.186$, $p = 0.057$), confirmant le fait que l'ESS mesure de manière subjective le même phénomène que le TILE (propension à l'endormissement, voir chapitre 5).

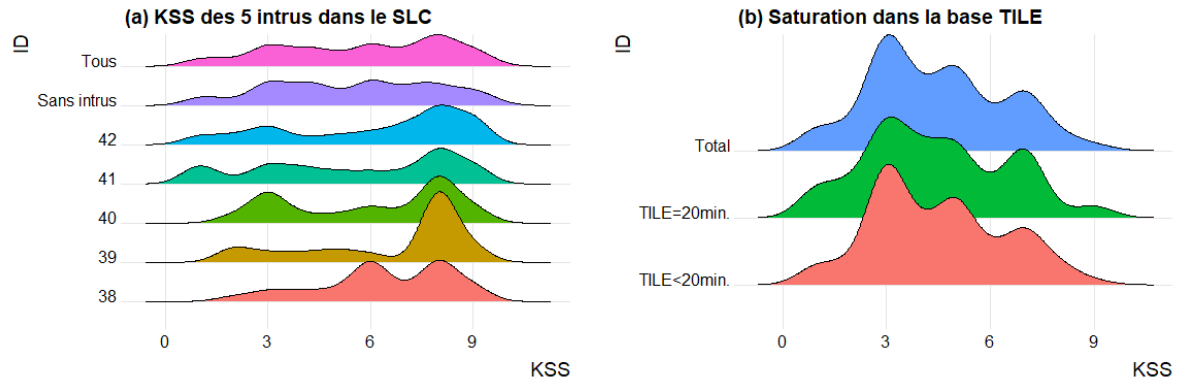

FIGURE 8.8 – Influence des locuteurs enregistrés un très grand nombre de fois sur la distribution de la KSS dans le SLC (gauche). Influence des saturations des latences d'endormissement sur la distribution des KSS dans la base TILE (droite).

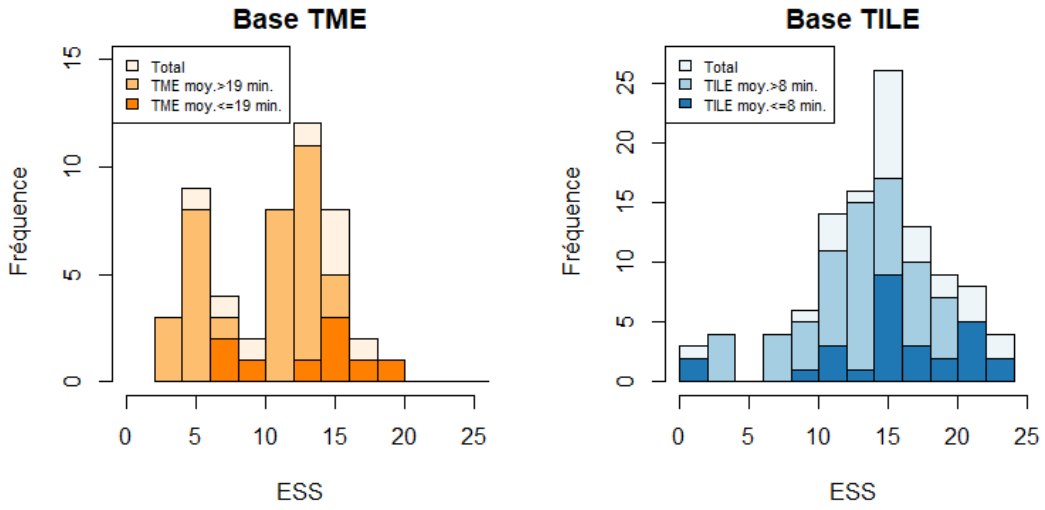

FIGURE 8.9 – Distributions de l'ESS sur la base TME (gauche) et la base TILE (droite).

8.3 Métadonnées

8.3.1 Sexe

Dans le SLC, la proportion entre femmes et hommes est presque équilibrée à l'échelle des locuteurs, mais un déséquilibre important existe à l'échelle des échantillons ($\chi^2, p < 2.2 \times 10^{-16}$). Ce déséquilibre est principalement dû aux cinq locuteurs enregistrés un très grand nombre de fois : sans ceux-ci, la proportion est presque équilibrée à l'échelle des échantillons (3084 échantillons enregistrés par des femmes, 2090 par des hommes) et leur répartition entre les deux classes de somnolence est quasiment équilibrée ($\chi^2, p = 0.16$).

Dans la base TME, les 57 locuteurs sont composés de 39 hommes et 18 femmes. La distribution des locuteurs dans les classes de somnolence par sexe semble déséquilibrée mais un test du χ^2 montre l'indépendance entre ces deux variables à la fois au niveau des locuteurs ($\chi^2, p = 0.15$) et des échantillons ($\chi^2, p = 0.056$).

Contrairement à la base TME, le déséquilibre de la répartition du sexe dans les classes de

somnolence n'est pas significatif à l'échelle des locuteurs dans la base TILE ($\chi^2, p = 0.16$) mais l'est à l'échelle des échantillons ($\chi^2, p = 9.6 \times 10^{-3}$).

8.3.2 Âge

Les locuteurs de la base TME (âge moyen : 46.3 ans) et ceux du TILE (âge moyen : 35.9 ans) sont plus âgés que ceux du SLC (âge moyen : 24.9 ans) et du SLEEP corpus (âge moyen : 27.6 ans). Les distributions des âges sur ces deux corpus sont représentées dans la figure 8.10. Dans les deux premiers corpus, il y a très peu de différence entre les patients S et NS (MW, $p = 0.93$ pour la base TME, $p = 0.96$ pour la base TILE). Leur différence avec le SLC et le SLEEP peut être expliquée à la fois par le fait que ces derniers contiennent des enregistrements de personnes mineures et par la population pathologique qui est enregistrée dans les bases TME et TILE, pour lesquels la prévalence des maladies est liée à l'âge.

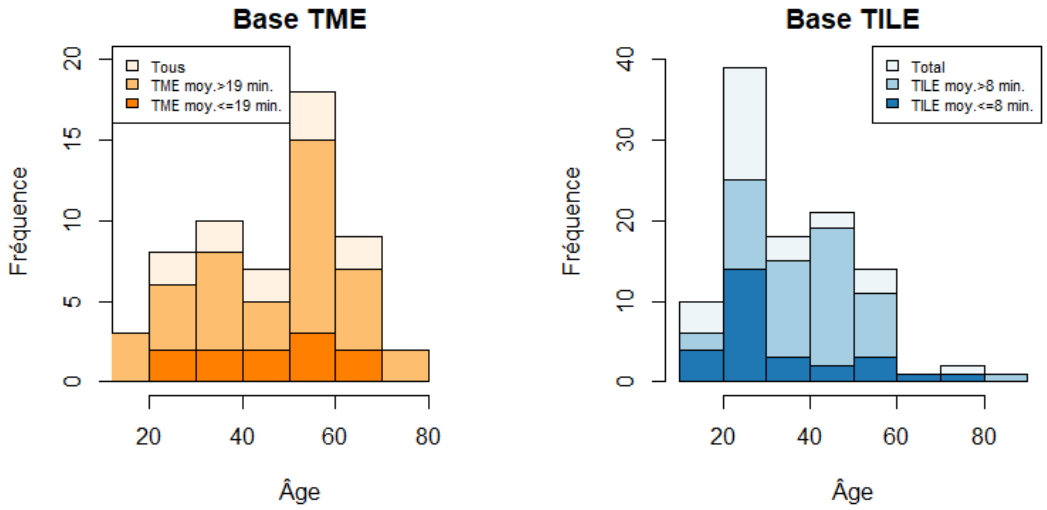

FIGURE 8.10 – Distribution de l'âge des locuteurs des bases TME (gauche) et TILE (droite).

8.3.3 Indice de Masse Corporelle – IMC

Les distributions de l'IMC des patients des bases TILE et TME sont représentées dans la figure 8.11. Le déséquilibre significatif du niveau d'IMC entre les deux classes de somnolence dans la base TME (test de Mann-Whitney, $p = 0.003$) peut être dû au déséquilibre entre sexes, mais pas seulement. En effet, dans la population française, l'IMC moyen est de 25.8kg/m^2 avec seulement une faible différence entre les hommes (25.8kg/m^2) et les femmes (25.7kg/m^2) (Verdot *et coll.*, 2017). Or, ces valeurs en population générale sont plus élevées que celles observées dans la base TME. Au contraire, dans la base TILE, l'IMC est plus proche des valeurs normatives précédentes, avec un IMC moyen de 24.2kg/m^2 et pas de différence significative entre les classes de somnolence (test de Mann-Whitney, $p = 0.89$).

8.3.4 Pathologies

Dans la base TME, le seul diagnostic disponible est celui pour le syndrome d'apnée obstructive du sommeil (SAOS), qui est équilibré entre les classes de somnolence ($\chi^2, p = 0.21$). Ce facteur est cependant à prendre en compte lors de l'estimation de la somnolence dans

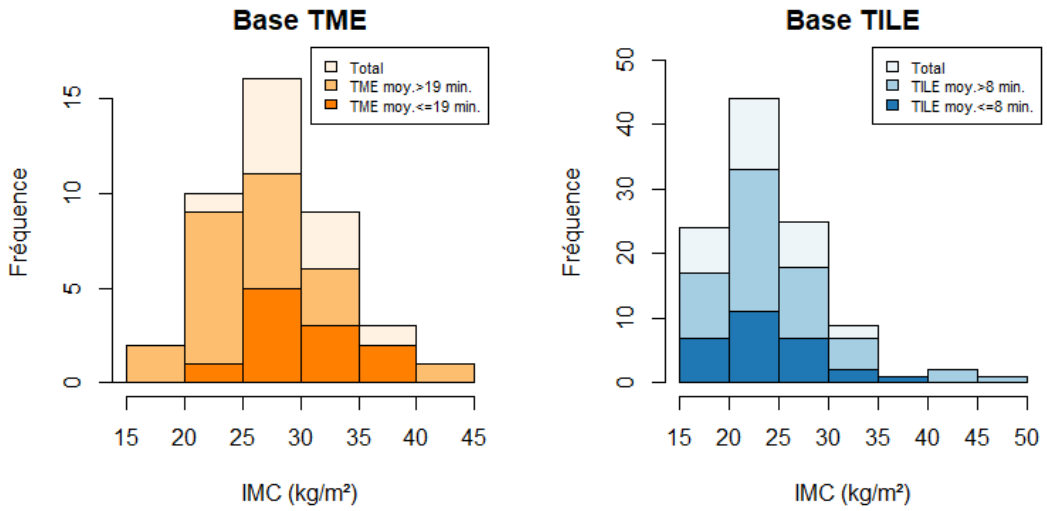

FIGURE 8.11 – Distribution des IMC dans la base TME (gauche) et la base TILE (droite).

la voix puisque ces patients peuvent présenter des variations de formes des voies aériennes supérieures.

Dans la base TILE, la distribution des latences moyennes d'endormissement en fonction des pathologies diagnostiquées est représentée dans la figure 8.12. En fusionnant les classes «Narcolepsie Type 1» et «Narcolepsie Type 2» dans une unique classe «Narcolepsie», les patients appartenant à cette catégorie ont une latence d'endormissement significativement plus faible que les autres sujets. Concernant les autres catégories, les autres sujets ont des latences d'endormissement équivalentes.

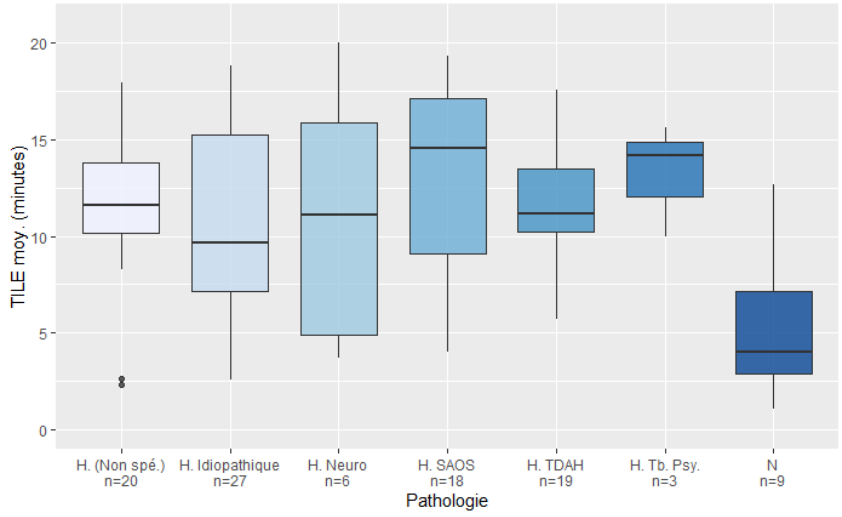

FIGURE 8.12 – Latence moyenne d'endormissement au TILE en fonction des diagnostics des patients. H. : Hypersomnie, SAOS : Syndrome d'Apnée Obstructive d'Apnée du Sommeil, TDAH : Trouble Déficitaire de l'Attention avec ou sans Hyperactivité, Tb. Psy. : Troubles psychiatriques, N : Narcolepsie (Type 1 ou 2).

Chapitre 9

Discussion et recommandations

Sommaire

Ce chapitre introduit, à travers la précédente comparaison entre les quatre principaux corpus pour la détection de la somnolence dans la voix, les questions que la conception d'une base de données pour la détection de la somnolence dans la voix peut soulever. Nous discutons les implications de ces choix, et nous proposons des recommandations globales concernant la conception d'un tel corpus.

Ces discussions, faites sur l'exemple de la somnolence, peuvent dans leur majorité être généralisées à tout corpus étudiant une pathologie ou un symptôme à travers la voix. De même, cette discussion est menée sur les corpus précédemment comparés et traitant de la somnolence. Dans le cas où certaines questions ne peuvent être tranchées uniquement avec les éléments inclus dans ces corpus, nous avons été chercher dans des tâches connexes (la détection de la dépression dans la voix par exemple) des éléments de réponses facilement transposables à notre problématique.

9.1 Choix des sujets

9.1.1 Sujets sains, population générale, patients

Lorsque l'on conçoit un tel corpus, une des premières questions qui apparaît est la population qui va être incluse. Concernant la détection de la somnolence ou de la SDE dans la voix, deux populations peuvent être envisagées : des patients lorsque le projet sous-jacent a pour but le suivi d'une telle population ; ou des personnes tirées de la population générale dans le cas où l'on souhaite travailler sur un état *général* de somnolence. Ainsi, dans les corpus TME et TILE, les sujets enregistrés sont des patients du service universitaire de médecine du sommeil du CHU de Bordeaux, tandis que dans le SLC et le corpus SLEEP, les sujets inclus sont recrutés dans la population générale. Ceux-ci ont malgré tout rempli un questionnaire relatif aux symptômes de pathologies du sommeil (PSQI) pour s'assurer qu'ils ne présentent pas de telles pathologies.

La différence entre sujets recrutés parmi la population générale et sujets sains est un point clé ici : un sujet peut être recruté dans la population générale et avoir une pathologie du sommeil. Cependant, la notion de *sujet sain* ne s'arrête pas au critère de jugement principal de l'étude : idéalement, les sujets sains doivent être diagnostiqués négatifs non seulement pour les problèmes de sommeil, mais aussi pour tous les cofacteurs qui peuvent influencer à la fois leur sommeil (anxiété, dépression, fatigue ...) et leur voix (anxiété, laryngite, pathologies du langage, pathologie des voies aériennes ...).

9.1.2 Équilibrer la base de données

Équilibrer la base de données est important à la fois pour des raisons éthiques (Cirillo *et coll.*, 2020) mais aussi pour s'assurer que les caractéristiques vocales ou les phénomènes mesurés sont indépendants des caractéristiques propres des sujets, comme par exemple l'âge, le sexe, l'IMC, etc. Cependant, équilibrer toutes ces caractéristiques est techniquement impossible. Pour évincer ces biais des bases de données, nous proposons deux directions. Tout d'abord, l'inclusion systématique de ces informations dans les bases de données, de façon à ce que les ingénieurs en charge de la conception des algorithmes d'apprentissage automatique puissent les prendre en compte.

Ensuite, nous différencions deux cas. Pour étudier un phénomène en population générale, la meilleure pratique est d'inclure un grand nombre de sujets, de façon à ce que ces biais soient randomisés. Au contraire, si l'objet de l'étude est de se focaliser sur une pathologie ou

un symptôme ayant une grande prévalence dans une population spécifique, se restreindre à ce type de population permet une bonne généralisation des concepts sous la contrainte de certaines hypothèses (Schnack et Kahn, 2016).

9.1.3 Niveau de lecture et des capacités de communication orale

Dans le cas d'une tâche de lecture, lire à voix haute est une tâche impliquant de nombreux processus neurolinguistiques et neuromoteurs (cf. chapitre 1). Ceux-ci peuvent être à la fois influencés par la somnolence et les capacités de lecture du locuteur. Pour s'assurer que toutes les informations extraites des enregistrements audio pour élaborer des algorithmes ne sont influencées que par le phénomène étudié (la somnolence dans notre cas) et non par les capacités de lecture du lecteur, les critères d'inclusion et d'exclusion doivent être adaptés afin de n'inclure que les patients ayant un niveau de lecture suffisant et d'exclure les patients ayant des pathologies du langage. Cette pratique a été introduite pour la détection de la somnolence dans la voix à travers une collaboration avec des orthophonistes, présentée dans le chapitre 7.

Les mêmes précautions doivent être prises pour des tâches de parole spontanée, afin de s'assurer que les hésitations et tous les autres marqueurs vocaux extraits sont liés exclusivement à la somnolence, et non à l'état émotionnel ou une pathologie.

Par exemple, les patients présentant une dysphonie, un trouble anxieux ou des problèmes de cordes vocales peuvent avoir une voix différente des autres sujets dans la base de données, qui n'est pas due à la somnolence, mais à des facteurs extérieurs.

9.1.4 Conclusions et recommandations pour le choix des sujets

- **Choix de la population**
  - Dans un contexte de suivi de patients, les sujets doivent être des patients avec des comorbidités contrôlées ;
  - Enregistrer des sujets sains semble plus adapté pour l'étude de la somnolence en population générale. Même si cela nécessite des critères d'inclusion et d'exclusion plus exigeants, ce qui pourrait rendre le recrutement plus difficile, cela permet d'éviter toute interférence entre le phénomène mesuré et les co-morbidités.
- **Caractéristiques des sujets**
  - Lorsque l'on travaille sur une hypothèse portant sur la population générale, le nombre de sujets doit être suffisant pour randomiser tous les cofacteurs ;
  - Au contraire, lorsque l'on travaille avec des patients, ils doivent avoir été diagnostiqués comme porteur de la pathologie ;
  - Dans les deux cas, les cofacteurs doivent être inclus dans les bases de données.
- **Troubles de l'élocution ou de la lecture**
  - Les sujets ne doivent pas être atteints d'une pathologie ou d'un trouble affectant leur capacité à lire un texte ou parler naturellement.

9.2 Conception des sessions d'enregistrement

Une fois les critères d'inclusion et d'exclusion fixés, de nombreuses configurations d'enregistrement sont possibles.

9.2.1 Égalité du nombre de sessions d'enregistrement par locuteur

Tout d'abord, la meilleure pratique lors de la construction d'une base de données est de s'assurer que le même nombre d'enregistrements est réalisé pour chaque locuteur, dans les mêmes conditions. Or, ceci n'est pas le cas dans le SLC, dans lequel certains locuteurs ont été enregistrés un nombre de fois très largement supérieur aux autres. Cela a pour conséquence des déséquilibres marqués concernant les distributions du sexe et des classes de somnolence dans ce corpus. De plus, le sexe et la KSS sont les seules données collectées dans le SLC, mais une surreprésentation de ces locuteurs pourrait également avoir créé des biais sur des critères qui ne sont pas mesurés dans ce corpus, comme l'âge, l'IMC, etc.

9.2.2 Nombre d'enregistrements par locuteur

Nous avons précédemment affirmé que la meilleure pratique est d'enregistrer chaque locuteur le même nombre de fois. Mais combien de fois est-il nécessaire d'enregistrer chaque sujet ? Un enregistrement vocal contient à la fois l'expression de trait du locuteur (âge, sexe, somnolence au long cours ...) et de son état (émotion, fatigue, somnolence, cycle circadien ...). Par conséquent, isoler un phénomène de l'autre nécessite soit de mesurer les deux et de prendre en compte le facteur indésirable lorsqu'on étudie la variable d'intérêt ; soit de randomiser le facteur indésirable. Les deux stratégies impliquent de multiples mesures.

Dans le TME et le TILE, les latences d'endormissement sont mesurées à différents moments standardisés avant d'être moyennées, pour estimer le trait de somnolence du patient, indépendamment de ses variations à court terme : l'état du locuteur varie au cours des enregistrements, alors que ses traits restent invariants. Pour ce qui est de l'estimation d'état à court terme, la même procédure a été appliquée dans la base TILE avec la collecte de la KSS : les mesures régulièrement espacées permettent d'estimer les traits du locuteur qui restent constants avant de considérer les variations par rapport à ceux-ci pour obtenir les états du locuteur. La collecte de la KSS à intervalles réguliers a également été effectuée dans une sous-partie du SLC, enregistrée dans un simulateur de conduite (Golz *et coll.*, 2007). Entre des sessions de 45 minutes de conduite simulée, les sujets réalisent diverses tâches vocales, et remplissent une KSS.

Cependant, excepté quand la collecte des données se greffe sur une autre étude, il semble compliqué d'enregistrer les sujets sur des périodes assez longues pour qu'ils soient dans différents états. Pour éviter la création de biais entre l'identité du locuteur et les labels, dus à la surreprésentation de certains locuteurs à travers des échantillons avec de faibles variations d'état, Huckvale *et coll.* (2020) suggèrent de concentrer les efforts dans l'enregistrement d'un nombre important de locuteurs, enregistrés un faible nombre de fois. Cette pratique permet de randomiser les traits du locuteur et permet une estimation correcte de l'état du locuteur, indépendamment de ses caractéristiques propres. Le contrecoup de cette méthode est qu'un nombre suffisant de sujets est nécessaire pour randomiser les caractéristiques des locuteurs, ce qui peut être difficile au regard des critères d'inclusion et d'exclusion de l'étude.

9.2.3 Durées des sessions d'enregistrement

Dans le corpus SLEEP, les sessions d'enregistrement ont une durée comprise entre 15 minutes et une heure, mais les échantillons ont une durée maximale de 5 secondes : les enregistrements ont été découpés pour augmenter le nombre d'échantillons dans le corpus. Cela permet notamment de tirer le maximum de la présence chaque locuteur en les enregistrant

durant une durée la plus longue possible, ou encore de créer un large corpus annoté, à partir de peu de sujets. Mais cela a aussi trois principaux inconvénients.

Tout d'abord, les questionnaires psychométriques de mesure de la somnolence qui sont utilisés pour annoter les données sont conçus pour une certaine durée de validité. Puisque l'état du locuteur peut varier de manière rapide, il est possible de faire plusieurs mesures du même locuteur dans différents états au sein d'une même session d'enregistrement. Cependant, la KSS n'a pas été conçue et validée pour des mesures répétées rapprochées (Kaida *et coll.*, 2006), mais pour des durées de l'ordre de la dizaine de minutes (l'énoncé précise « dans les 10 dernières minutes »). Réitérer le remplissage d'une KSS après moins de 10 minutes pourrait donner des scores différents, non en raison d'un changement de somnolence du patient, mais parce que le questionnaire n'a pas été conçu pour. Un remplissage de KSS par intervalle de 10 minutes devrait être suffisant pour estimer la somnolence du locuteur sur cet intervalle et collecter assez de données par session.

Ensuite, le nombre d'échantillons pourrait ne pas être le même pour tous les locuteurs, surreprésentant certains locuteurs et les biais liés à chaque session d'enregistrement. De plus, enregistrer chaque patient un très grand nombre de fois va dans la direction opposée de la discussion du paragraphe précédent.

Enfin, des sessions d'enregistrement de cette longueur induisent une fatigue vocale (Caraty et Montacié, 2014) et cognitive qui peut affecter à la fois la production vocale et l'estimation de la somnolence. Si découper les échantillons est possible – avec quelques précautions, comme par exemple la taille minimale des échantillons, discutée dans la section 9.4 – les sessions d'enregistrement doivent garder une durée raisonnable et équilibrée entre les participants.

9.2.4 Lieu d'enregistrement

Après avoir discuté de la conception des sessions d'enregistrement, la question du lieu d'enregistrement nécessite d'être élucidé. La littérature mentionne trois différentes configurations : dans une chambre d'hôpital (bases TILE et TME), dans un simulateur de conduite (SLC) et dans une salle de lecture (SLEEP).

Deux choix méthodologiques proposés ici s'opposent. D'une part, les enregistrements peuvent être réalisés des conditions aussi proches que possible du but final de l'étude, comme par exemple les simulateurs de conduite. Cependant, cet environnement d'enregistrement crée un biais dans le comportement et l'auto-évaluation du sujet, limitant l'exploitation des résultats ainsi produits.

De l'autre, les installations logistiques déjà présentes peuvent être en faveur des enregistrements en conditions hospitalières. Celles-ci peuvent certes être stressantes pour les sujets qui n'y sont pas habitués (Aydin Sayilan *et coll.*, 2020), mais elles offrent un accès facilité aux équipements d'enregistrement PSG, seul moyen de mesure "objectif" de la somnolence, et des conditions parfaitement contrôlées : les participants sont traités de manière équivalente, ont leurs repas aux mêmes heures, et leur nuit de sommeil avant le test médical est mesurée et contrôlée. Tous ces avantages tendent à encourager les enregistrements dans des conditions hospitalières dans un premier temps, avant d'étendre le problème à des conditions écologiques.

9.2.5 Qualité d'enregistrement

La qualité audio des enregistrements peut être affectée par plusieurs sources de bruits à la fois dans des conditions écologiques (le bruit du simulateur de conduite ou du trafic routier dans le cas de la conduite en conditions réelles par exemple) et des conditions hospitalières

(le bruit de la ventilation par exemple). Par conséquent, un lieu calme et non réverbérant doit être favorisé.

Un autre aspect impactant la qualité des enregistrements est le matériel utilisé, qui prolonge la précédente dichotomie entre réalisme et contrôle de l'environnement de la collecte des données. Alors qu'un nombre croissant d'études se basent sur des enregistrements effectués avec des smartphones (par exemple (Huang *et coll.*, 2018)), une des craintes liées à l'utilisation de ces enregistrements repose sur une possible dégradation des marqueurs acoustiques qui en sont extraits. Une récente étude n'a cependant pas trouvé de différence de performances sur la détection de la maladie de Parkinson à partir de la voix, en comparant des enregistrements de haute qualité avec des données récoltées par smartphones (Vasquez-Correa *et coll.*, 2021). De plus, les marqueurs extraits d'échantillons vocaux ne se limitent pas à des paramètres acoustiques : divers marqueurs linguistiques sont pertinents pour la détection de diverses pathologies (Aloshban *et coll.*, 2021; Martin *et coll.*, 2020a) et sont moins influencés par la qualité d'enregistrement que les paramètres acoustiques.

9.2.6 Conclusion et recommandations sur les sessions d'enregistrement

- **Multiplier les mesures** Nous encourageons
  - Soit de multiplier les mesures du même locuteur dans différents états pour l'estimation de trait du locuteur ;
  - soit de multiplier le nombre de sujets enregistrés pour randomiser les traits des locuteurs pour l'estimation de traits ;
- **Durée des sessions d'enregistrement**
  - À la fois pour réduire les biais et la fatigue causée par la tâche, nous encourageons la standardisation et la limitation de la durée des sessions d'enregistrement ;
- **Lieu d'enregistrement**
  - Le milieu hospitalier semble être un choix pertinent dans un premier temps, avant de passer à des conditions aussi proches que possible de l'application finale désirée ;
- **Qualité des enregistrements**
  - Nous encourageons, lorsque c'est possible, d'enregistrer les sujets dans un environnement calme et non réverbérant ;
  - Le choix entre un smartphone et un microphone de haute qualité dépend des marqueurs extraits ensuite des fichiers audio. Cependant, s'il est possible de dégrader des enregistrements de bonne qualité pour les faire imiter des enregistrements de smartphones, l'inverse n'est pas possible.

9.3 Tâche vocale

9.3.1 Parole spontanée vs Lecture

Que cela soit lors de l'interaction avec un médecin physique ou virtuel, ou lors d'un appel téléphonique, la parole spontanée peut être facilement enregistrée en conditions écologiques. Par conséquent, concevoir un corpus basé sur la parole spontanée des sujets semble la façon la plus naturelle d'enregistrer la voix des participants. Cependant, ce paradigme souffre de trois désavantages.

Tout d'abord, lorsque l'on élabore des algorithmes d'apprentissage automatique, la voix des sujets doit être polluée le moins possible par ses émotions, pour éviter aux systèmes de discriminer l'état émotionnel des sujets à la place de leur niveau de somnolence. Les questions

ouvertes impliquant la mémoire ou les émotions induites par l'observation et la description d'une peinture – comme c'est le cas dans le corpus SLEEP – changent l'état du locuteur. Deuxièmement, comme étudié précédemment par Stasak *et coll.* (2018), ce paradigme souffre d'un autre désavantage : si les questions sont identiques, les échantillons n'ont ni le même contenu vocalique, ni les mêmes tailles. Cela empêche toute comparaison pertinente entre les échantillons et crée des biais entre eux. Finalement, ce paradigme ne garantit pas de taille minimale des échantillons : durant une étude préliminaire au corpus TME basée sur des tâches de parole spontanée, certains patients répondaient de la manière la plus courte et la plus concise possible, voire ne répondaient pas du tout, ce qui se traduit par des échantillons trop courts pour être utilisables.

D'un autre côté, la parole spontanée bénéficie de sa proximité avec les conditions écologiques et de la possibilité d'analyser le contenu de ce qui est dit. En effet, une récente étude sur la détection de la dépression a fait le lien entre une augmentation de mots à valence négative dans le langage et l'humeur (*mood*) des patients atteints de troubles bipolaires, permettant ainsi de les détecter grâce à la voix (Matton *et coll.*, 2019). Un tel système pourrait être développé pour la somnolence et la somnolence excessive avec la détection de mots tels que « somnolent », « fatigué », « crevé », ...

Les tâches de lecture pour leur part permettent d'assurer que les enregistrements ont des tailles similaires, le même contenu et qu'elles sont moins polluées par les émotions. Même si cette tâche semble difficile à implémenter en conditions écologiques, cela permet d'étudier la voix des sujets dans des conditions parfaitement contrôlées et avec des contenus comparables, avant d'étendre les études à des tâches vocales plus naturelles. De plus, la lecture de textes permet une comparaison à une référence, pour concevoir de nouveaux biomarqueurs comme par exemple les erreurs de lecture ou les erreurs faites par les systèmes de reconnaissance automatique de la parole (cf. chapitre 13 et 14).

9.3.2 Performances en fonction de la tâche vocale dans la classification de la dépression

À notre connaissance, aucune étude comparative entre les différents types de tâches vocales sur la somnolence n'a été menée. Cependant, dans le domaine de la détection de la dépression dans la voix, une étude a comparé de la parole lue (*Grandfather Passage*) et spontanée (sujets expliquant comment ils se sentent physiquement et émotionnellement) pour estimer la dépression chez des patients sous traitements (Espy-Wilson *et coll.*, 2019). Cette étude affiche un taux de classification correcte de 64.3% en utilisant la parole lue, contre 71.4% sur la parole spontanée, ce qui semble encourager l'utilisation de la parole spontanée lors de la conception d'un corpus.

Au contraire, une autre étude menée par Kiss et Vicsi (2017), basée sur la fable *La bise et le soleil*, et sur une interview durant entre 5 et 10 minutes pour la parole spontanée, rapporte des performances identiques pour les deux modalités (respectivement 83% et 85%), encourageant à parts égales les deux choix.

De plus, ces deux modalités sont détectées avec des marqueurs vocaux différents : la dépression à travers la parole spontanée est détectée avec des marqueurs de prosodie, tandis que pour la lecture les marqueurs pertinents sont des marqueurs extraits des formants.

9.3.3 Choix du texte

Cette section discute des contraintes qu'un texte doit satisfaire pour permettre la détection de la somnolence dans la parole.

Tout d'abord, les textes choisis doivent être différents pour chaque session d'enregistrement. En effet, dans le cas d'un TME ou d'un TILE, si le texte est le même au cours des itérations du test, des effets d'apprentissage et de fatigue apparaîtraient en raison de la répétition. Cependant, les textes doivent être les plus similaires possibles, pour éviter tout biais dû au texte en lui-même. Un des points communs qu'ils doivent partager est leur taille : s'ils ont des tailles différentes, comme c'est le cas dans la base TME, cela introduit des biais à la fois de fatigue et de temps de maintien de l'attention nécessaire à la lecture des textes, et vis-à-vis des marqueurs.

Deuxièmement, dans la base TME, le contenu est complètement différent d'une itération à l'autre, créant un biais d'émotions qui peut ensuite polluer les enregistrements : une personne ne ressent pas les mêmes émotions en lisant un résumé d'article de vulgarisation scientifique ou une fable.

De plus, le contenu ne doit pas interférer avec la mesure de la somnolence : un texte trop stimulant ou ennuyeux pourrait changer le niveau de somnolence du locuteur, ainsi que ses mesures (KSS, TILE, TME ou autre). En conséquence, les textes choisis doivent avoir un contenu qui induit le même état émotionnel pour toutes les lectures, à la fois pour la production vocale que pour la mesure médicale de la somnolence.

À part la taille et le contenu émotionnel des textes, d'autres paramètres sont à prendre en compte. Par exemple, les deux fables du corpus TME et les cinq textes du corpus TILE ne sont pas exactement équivalents puisque, dans les deux cas, certains contiennent des dialogues et d'autres non : puisque les dialogues nécessitent un autre niveau d'attention visuoattentionnelle, les textes représentent des niveaux de difficulté différents. Pour éviter de telles disparités, nous recommandons de choisir des textes sans dialogues pour assurer que tous les patients sont considérés équitablement concernant ce point. De même, le corpus TILE est basé sur des extraits du Petit Prince, en raison de son vocabulaire et de sa grammaire qui sont simples, pour éviter l'interférence du niveau de lecture du locuteur avec la détection de la somnolence. Cette pratique devrait être généralisée à tout corpus basé sur de la lecture de textes.

Pour conclure, nous ne voyons pas d'intérêt particulier à ajouter un défi supplémentaire en proposant aux patients des textes qui ne sont pas dans leur langue natale, comme cela a été proposé dans le corpus SLEEP.

9.3.4 Conclusion et recommandations sur le choix de la tâche vocale

- **Tâche vocale**
  - Les tâches de lecture permettent de contrôler le contenu des enregistrements, mais sont éloignées des conditions écologiques ;
  - Les tâches de parole spontanée conduisent à des enregistrements de tailles très diverses, mais permettent l'analyse du contenu du discours.
- **Contenu des textes**
  - Les textes doivent être les plus similaires les uns des autres en termes d'émotions inférées, de phonétique, avec une grammaire et un vocabulaire simples.

9.4 Durée des échantillons audio

Cette section propose une étude des durées minimales et maximales des échantillons audio enregistrés. Connaître les durées minimale et maximale théoriques des échantillons audio permet à la fois de concevoir la tâche vocale (longueur du texte ou durée minimale de la réponse à une tâche de parole spontanée), et de choisir la taille minimale des tronçons si l'on souhaite découper les enregistrements pour augmenter le nombre d'échantillons.

Dans le corpus SLEEP, tous les échantillons ont une durée inférieure à 5 secondes, avec une durée moyenne de 3.87s. En effet, ces enregistrements proviennent d'échantillons audio qui ont été ensuite découpés en tronçons d'environ 4 secondes. La même pratique est couramment employée pour la détection de la dépression (par exemple dans (Ma *et coll.*, 2016; Vasquez-Correa *et coll.*, 2020)), pour laquelle cette durée de 4 secondes a été démontrée comme optimale pour le corpus employé. Une autre étude sur la même tâche utilise des tronçons de 10 secondes (Nasir *et coll.*, 2016), mais l'objectif derrière ce choix n'est pas clairement exposé.

Cependant, ces résultats doivent être questionnés : la détection de la somnolence dans la voix est une tâche différente, et la somnolence peut s'exprimer différemment dans la voix : la question de la taille des échantillons permettant la détection de la somnolence est encore à élucider.

9.4.1 Première approche – Performances dans le SLC

Une première approche proposée dans (Martin *et coll.*, 2019) a étudié les performances d'un classifieur entraîné sur un sous-corpus du SLC composé uniquement des tâches de lecture, décrit dans le chapitre 11. Une fois les marqueurs vocaux et les hyperparamètres déterminés sur les sous-corpus d'entraînement et de développement, nous avons entraîné le système et calculé les performances sur les sous-corpus d'entraînement + développement vs test avec différents seuils de durée minimale pour les échantillons.

Le résultat de cette expérience est présenté dans la figure 9.1. Un premier palier de performances est atteint à 71% d'UAR (cf. chapitre 16 pour plus d'information) en ne sélectionnant que les échantillons de plus de 4 secondes. Un deuxième palier, moins franc en raison du faible nombre d'échantillons dont la durée est supérieure à 7 secondes, semble se profiler à partir de 8 secondes. C'est donc cette valeur que nous avons dans un premier temps retenue comme longueur minimale pertinente pour la détection de la somnolence dans la voix.

9.4.2 Deuxième approche – Convergence des marqueurs acoustiques de la voix

Une deuxième approche que nous avons proposée dans (Martin *et coll.*, 2020c), est basée sur les marqueurs acoustiques extraits des échantillons audio.

Méthode

Nous avons découpé tous les échantillons audio des corpus SLC et TILE en tronçons contenant uniquement la première seconde de l'échantillon, uniquement les deux premières secondes de l'échantillon, uniquement les trois premières secondes de l'échantillon, etc. On obtient ainsi des échantillons de taille croissante, sur lesquels on calcule les marqueurs acoustiques présentés dans le chapitre 11. Pour éviter un biais qui serait propre à nos marqueurs, nous extrayons également les marqueurs proposés lors de la conférence Interspeech 2011 grâce à la boîte à outils `openSMILE` (Eyben et Schuller, 2015) pour comparaison.

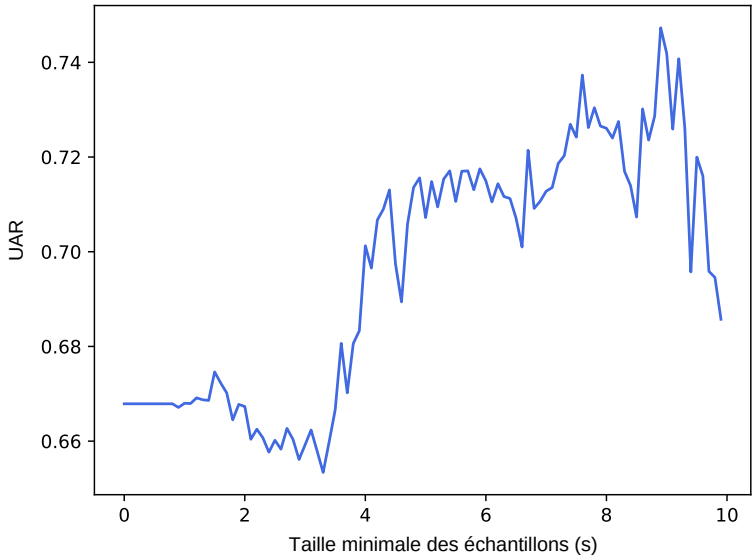

FIGURE 9.1 – Performances (UAR) du système (e) présenté dans le chapitre 11 en fonction de la taille minimale des échantillons sur les tâches de lecture du SLC.

Ensuite, nous calculons pour chaque échantillon la similarité cosinus entre le marqueur correspondant au tronçon de taille i secondes et celui correspondant au tronçon de taille $i+1$ secondes, issus du même fichier audio :

$$s_{i,i+1} = \frac{|X_i| \cdot |X_{i+1}|}{||X_i|| \cdot ||X_{i+1}||}$$

Ainsi, quand $s_{i,i+1}$ est proche de 1, X_i est proche de X_{i+1} : l'information supplémentaire apportée par la seconde supplémentaire entre les échantillons i et $i+1$ est faible. Nous calculons la moyenne et l'écart-type des $s_{i,i+1}$ pour tous les échantillons, et nous obtenons le graphe représenté dans la figure 9.2. Il représente l'information supplémentaire apportée par chaque seconde supplémentaire dans l'échantillon, à partir d'un échantillon contenant une seule seconde d'enregistrement.

Résultat

Une première remarque concerne la différence de valeurs entre l'évolution des marqueurs personnalisés et ceux extraits avec `openSMILE`. En effet, les valeurs de moyenne et d'écart-type de $s_{i,i+1}$ sont très proches respectivement de 1 et de 0 pour les marqueurs extraits avec `openSMILE`, et ce, quel que soit i. Nous faisons l'hypothèse que cela provient de la différence de taille des ensembles de marqueurs. En effet, dans le cas des marqueurs openSMILE IS11, une différence franche sur un nombre réduit de marqueurs aura peu d'impact sur la similarité cosinus calculée sur les 4 368 marqueurs, contrairement aux marqueurs personnalisés, au nombre de 44. Cependant, cette différence ne change pas l'interprétation faite de l'évolution de $s_{i,i+1}$. En effet, quel que soit l'ensemble de marqueurs considéré ou le corpus, pour une durée d'environ 8 secondes, la moyenne et l'écart-type de $s_{i,i+1}$ commencent à devenir stationnaires : toute information audio supplémentaire n'apporte que peu d'information vis-à-vis des marqueurs audio calculés sur une durée plus courte. Cette durée minimale concordant

avec les résultats de la précédente section, cette limite semble robuste pour la détection de la somnolence dans la voix grâce à des marqueurs acoustiques.

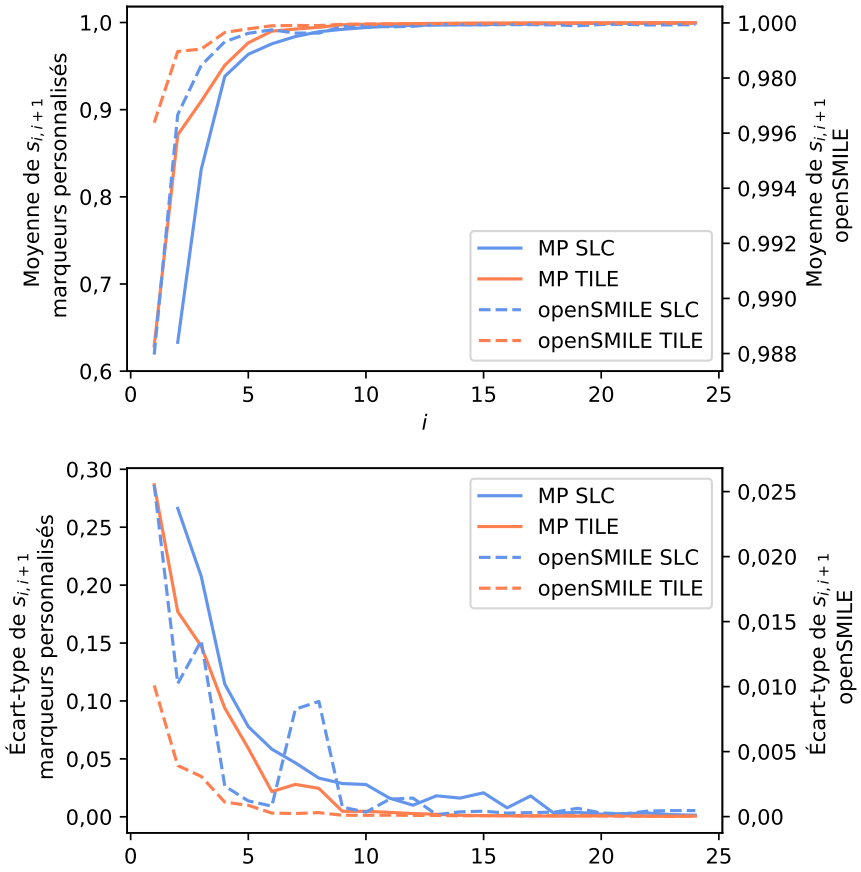

FIGURE 9.2 – Moyenne et écart-type sur les échantillons des distances cosinus entre les marqueurs acoustiques extraits d'un échantillon d'une durée de i seconde et ceux d'une durée de taille $i+1$ secondes. MP : Marqueurs personnalisés.

9.4.3 Troisième approche – base TILE et SLC

Une troisième approche, également basée sur les marqueurs vocaux calculés sur les corpus SLC et TILE, est proposée dans (Martin *et coll.*, 2021).

Méthode La figure 9.3 résume la méthode employée. Les premières étapes sont similaires à l'approche précédente (découpe des échantillons audio en tronçons de taille croissante et calcul des marqueurs acoustiques sur ceux-ci), la différence résultant dans la métrique employée pour étudier la convergence des marqueurs.

Dans le paragraphe précédent, le jugement s'effectue sur la stationnarité de la distance cosinus entre X_n et X_{n+1}, les marqueurs respectivement associés aux échantillons de taille n et $n+1$. Le principal inconvénient de cette métrique est sa sensibilité à la disparité des échelles de mesure : les changements de $s_{n,n+1}$ sont principalement dirigés par les changements des grandes valeurs (comme par exemple les fréquences, qui ont plusieurs ordres de grandeur de plus que les ratios associés).

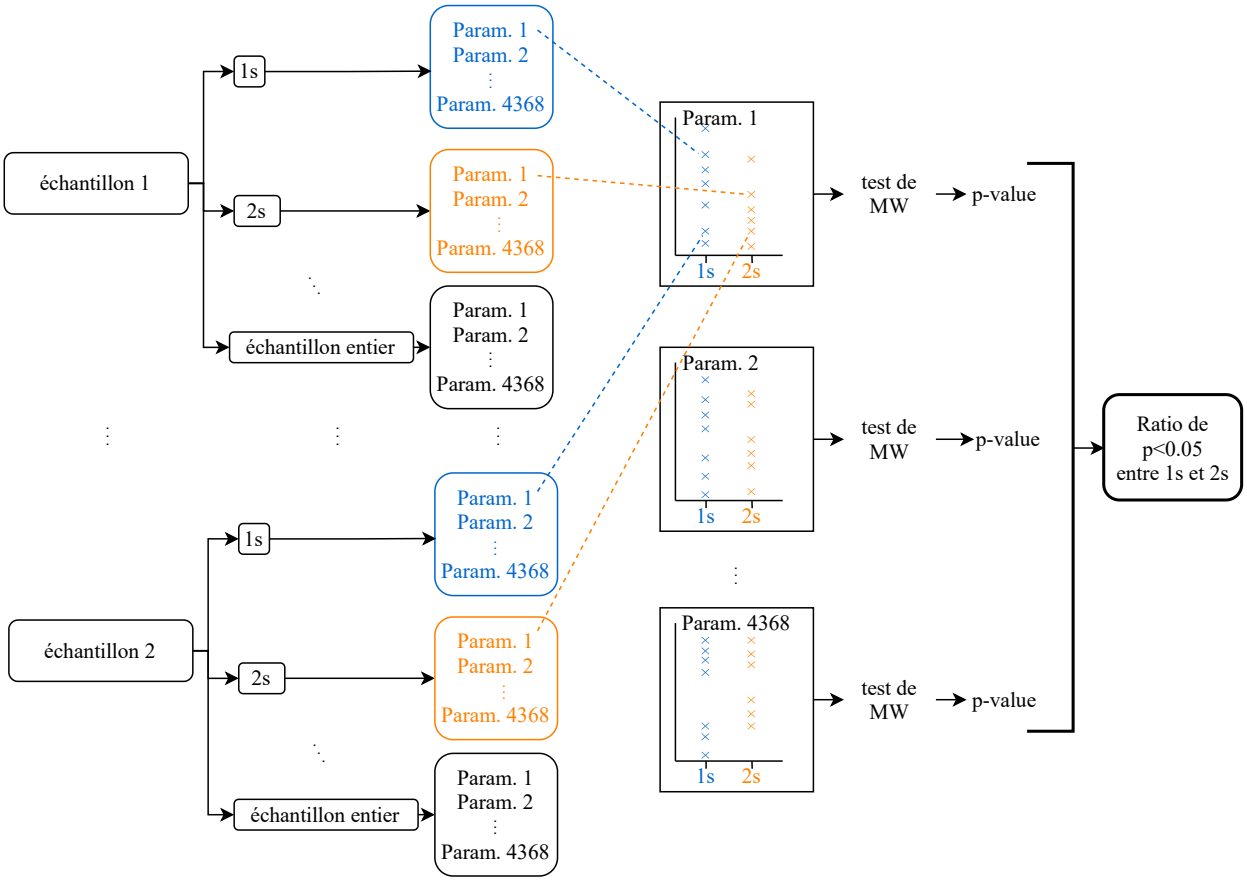

FIGURE 9.3 – Schéma explicatif de la méthodologie de la troisième méthode, basée sur la significativité de la différence entre les marqueurs extraits des échantillons de taille n secondes et ceux de taille $n+1$.

Dans cette approche, nous nous intéressons à la proportion de marqueurs pour lesquels la distribution de X_n est significativement différente de X_{n+1} (test de Mann-Whitney, $p < 0.05$). En effet, quand ces deux distributions sont significativement différentes, ajouter une seconde au fichier audio induit une variation significative des marqueurs, signifiant que le marqueur n'a pas encore convergé : augmenter la taille des échantillons apporte encore de nouvelles informations.

Résultats En appliquant le processus décrit précédemment aux quatre corpus, nous obtenons la figure 9.4. Puisque dans le corpus SLEEP les échantillons durent moins que 5 secondes, le graphique correspondant s'arrêterait au bout de 5 secondes avec plus de 75% des marqueurs IS11 n'ayant pas convergé. En conséquence, et pour favoriser la lisibilité de la figure, nous ne l'avons pas inclus dans la figure 9.4.

Pour ce qui est des trois autres corpus, une durée minimale de 20 secondes conduit à des ratios de paramètres n'ayant pas convergé inférieurs à 15%. Selon cette métrique, la précédente limite de 8 secondes conduit à plus de 25% de marqueurs acoustiques n'ayant pas convergé pour le SLC et le corpus TILE, et plus de 15% pour la base TME. Ainsi, une durée minimale de 8 secondes ne semble pas suffisante pour garantir la convergence des marqueurs. En conséquence, nous recommandons une taille minimale de 20 secondes pour les échantillons audio. Dans tous les cas, nous recommandons le test de la compatibilité entre les marqueurs employés et le contenu du corpus utilisé.

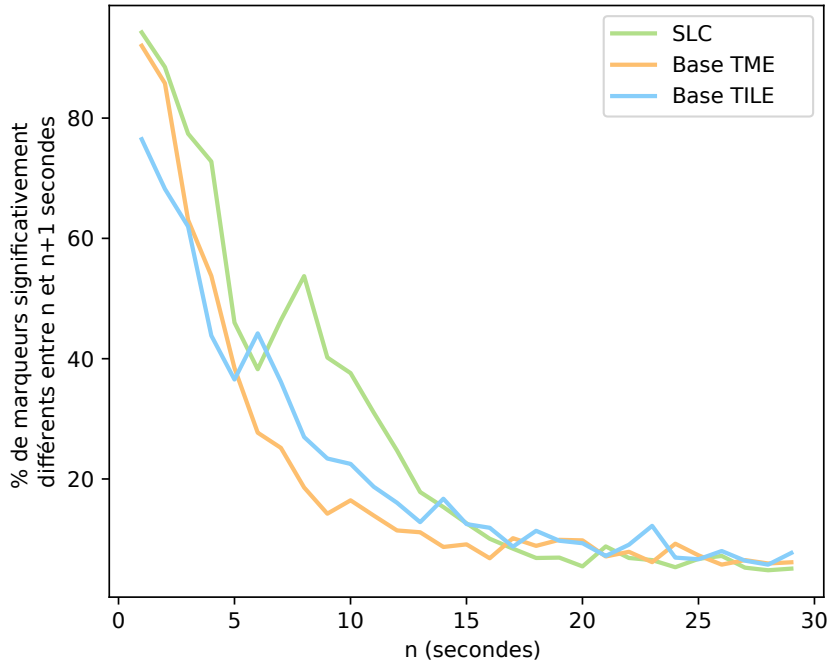

FIGURE 9.4 – Ratio de marqueurs significativement différents entre n et $n+1$ secondes sur les trois corpus SLC, TILE et TME, pour les marqueurs `openSMILE` IS11.

En ayant conscience que la convergence des marqueurs acoustiques de la voix n'est pas liée directement à la somnolence, travailler avec des marqueurs n'ayant pas convergé et dont la valeur aurait été différente si l'échantillon audio avait été plus long d'une seconde est périlleux. En effet, nous avons observé les mêmes résultats en utilisant les autres ensembles de marqueurs acoustiques proposés dans `openSMILE`, c'est-à-dire IS09 pour la reconnaissance d'émotions, IS10 pour la classification de la dépression, IS12 pour l'estimation de traits de locuteurs, et les marqueurs ComParE IS13. En conséquence, indépendamment de la tâche et des marqueurs acoustiques qui y sont habituellement associés, une taille minimale semble requise pour que les marqueurs convergent. De plus, tout échantillon plus long que la limite minimale conduira aux mêmes marqueurs que ceux qui ont la durée minimale proposée : en l'absence de consensus sur la question, nous recommandons de travailler avec des échantillons de 20 secondes minimum.

9.4.4 Taille des échantillons pour la détection de la dépression

Un autre argument en faveur de cette limite est le résultat obtenu dans une étude dédiée à la taille des échantillons pour la classification de la dépression dans la parole spontanée (Rutowski *et coll.*, 2019). En étudiant l'influence de la taille des échantillons (en nombre de mots) sur les performances de deux classifieurs basés sur des réseaux de neurones profonds, les auteurs concluent que 1) la taille minimale pour permettre une généralisation du concept est de 30-50 mots (environ 20 secondes dans le corpus utilisé), et 2) les performances de classification augmentent avec la taille des échantillons.

Un travail similaire est proposé dans (Di *et coll.*, 2021) a été mené sur des enregistrements

d'entretiens cliniques analysés à l'aide de i-vecteurs, pour lequel les auteurs concluent qu'une durée minimale de 40 secondes est nécessaire pour que les descripteurs audio atteignent leur palier de stabilité.

9.4.5 Taille maximale

Néanmoins, il faut aussi éviter les tâches vocales trop longues : certains sujets sont facilement ennuyés, et ces tâches pourraient induire l'expression de fatigue ou d'irritation dans la voix, biaisant la manifestation de la somnolence dans celle-ci. De plus, l'article précédemment cité (Rutowski *et coll.*, 2019) a montré que les performances des classifieurs étudiés saturent pour des échantillons dont la longueur est supérieure à 120-200 mots. En conséquence, nous recommandons de se restreindre à des longueurs raisonnables, entre une et deux minutes.

9.4.6 Conclusion et recommandations concernant la durée des échantillons

— **Durée des échantillons**
  — Une durée minimum de 20 secondes semble nécessaire pour observer la convergence des marqueurs acoustiques de la voix : nous recommandons de ne pas découper les enregistrements en tronçons d'une durée inférieure ;
  — Une durée maximale de 1 à 2 minutes est suffisante pour avoir un contenu pertinent à étudier, sans induire de fatigue ou d'ennui de la part du locuteur.

9.5 Annotation de la somnolence

Après avoir choisi la tâche vocale et conçu les sessions d'enregistrements, une question cruciale reste sans réponse : comment annoter les données ? Quand on travaille sur des processus neuropsychiatriques qui ne sont pas directement mesurables (la *somnolence*, la *fatigue*, les *émotions*) et dont les paradigmes sont ardus à formuler, cette question n'est pas triviale (Starke *et coll.*, 2020).

Formulation du problème

Dans le cas spécifique de la somnolence, une façon de déterminer la façon la plus pertinente d'annoter les données est de se demander :

1. Quel est le phénomène que l'on souhaite exactement mesurer ? La somnolence à court terme ? Une maladie du sommeil ? La fatigue ? Les performances au volant ? Toutes ces tâches sont parfois confondues [1] mais sont intrinsèquement différentes ;
2. Quelle est la population cible ? En effet, les tests et les questionnaires sont calibrés et validés sur des populations spécifiques et sont sensibles à la population à laquelle ils sont proposés.

Choisir la meilleure mesure possible pour détecter la somnolence dans la voix nécessite une collaboration étroite entre médecins et ingénieurs (Littmann *et coll.*, 2020). À partir des mesures de la somnolence utilisées dans les corpus précédemment présentés, nous avons identifié trois axes qui nous semblent définir un espace de validité des mesures de la somnolence pour l'annotation d'une base de données :

1. "Here, drowsiness is used as synonymous to sleepiness and fatigue, and antonymous to alertness" (Sparrow *et coll.*, 2019)

— l'adéquation entre la mesure et l'objectif (qu'est-ce que l'on désire mesurer ?) ;
— le respect de la nécessité pour les algorithmes d'apprentissage automatique d'avoir des données équilibrées (est-ce que les classes d'annotations sont différentes en raison de la mesure choisie ?) ;
— la conservation d'un sens médical (validité médicale).

Satisfaire ces trois contraintes est un but idéal. Lorsque l'on choisit une mesure de la somnolence, des compromis doivent être faits : l'important est d'être conscient des hypothèses qu'ils induisent. Le positionnement des mesures de la somnolence utilisées dans les 4 corpus présentés précédemment selon les trois axes proposés est représenté dans la figure 9.5.

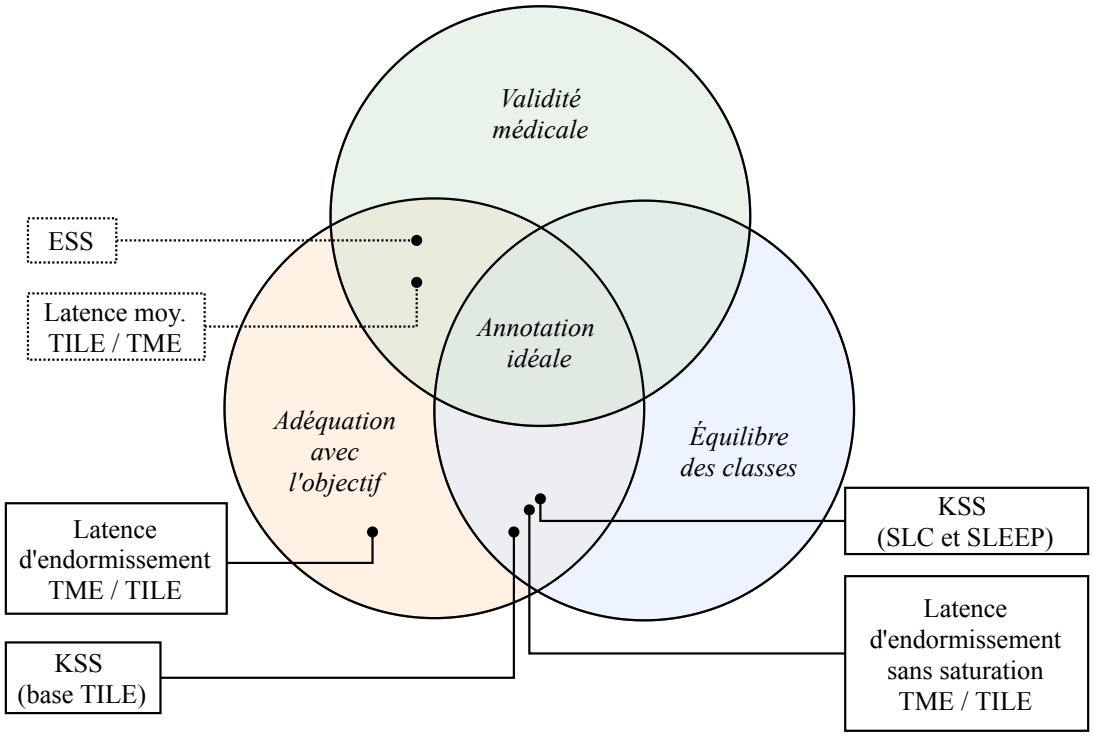

FIGURE 9.5 – Positionnement des mesures utilisées dans les corpus décrits dans ce chapitre selon les trois axes définissant un espace de validité pour l'annotation d'une base de données. Lignes pleines : mesures à court terme. Lignes pointillées : mesures au long cours.

9.5.1 Mesures de la somnolence utilisées dans les corpus présentés

Le SLC et le corpus SLEEP ont été conçus pour la détection de la somnolence à court terme chez des sujets sains. Le choix a été fait par les concepteurs de ces corpus d'utiliser un questionnaire médicalement validé sur cette population (la KSS), annoté à la fois par le sujet lui-même (autoquestionnaire), mais aussi deux annotateurs externes (hétéroquestionnaire). Cette mesure conduit à des classes relativement équilibrées (cf. figure 8.7), et semble adaptée au but de l'expérience. Le principal inconvénient de cette mesure est que la KSS sous cette forme n'est pas validée médicalement, soulevant une question cruciale pour la compréhension des mécanismes liant la voix à la somnolence : que mesure exactement ce score ? Tant que cette question reste sans réponse, réussir à atteindre des taux de classification correcte élevés peut certes mener à de nouvelles perspectives technologiques, mais l'outil n'aura un sens scientifique qu'une fois que la mesure de la somnolence utilisée sera validée et confrontée à

d'autres mesures de référence.

Dans la base TILE, la KSS a pour but de mesurer la somnolence subjective des patients atteints de diverses maladies du sommeil. Cette mesure correspond à l'objectif de détection de la somnolence à court terme, mais souffre de son absence de validité sur des populations pathologiques (Sangal, 1999; Ihler *et coll.*, 2020; Evangelista *et coll.*, 2020). La fiabilité des autonotations des locuteurs est donc remise en cause.

Les corpus TME et TILE ont été conçus avec des rôles différents du SLC et du SLEEP. Basés sur des tests médicaux validés, ils ont pour but le suivi de patients souffrants de maladies du sommeil. Les latences moyennes d'endormissement au TME et au TILE sont toutes deux des mesures neurophysiologiques de références de la somnolence excessive (Arand *et coll.*, 2005). De cette manière, ils remplissent le critère de validité médicale. Leur principal inconvénient réside dans les efforts nécessaires pour obtenir peu de données. En effet, la collecte des données pour un seul locuteur nécessite plusieurs enregistrements tout au long de la journée, et le nombre de locuteurs dépend directement du nombre de chambres hospitalières dédiées à ces tests. De plus, comme étudié dans la section précédente, les latences d'endormissement sont sujettes à des saturations, rendant leur exploitation par des algorithmes d'apprentissage automatique difficile.

Avant de les moyenner, les latences d'endormissement au TME et au TILE peuvent être des mesures respectives des pertes de vigilance et de propension à l'endormissement à court terme. Ces mesures souffrent cependant de deux inconvénients : d'une part, les latences d'endormissement individuelles ne sont pas des mesures de la somnolence instantanée validées médicalement ; d'autre part, elles souffrent elles aussi de phénomènes de saturation.

Collectée dans les corpus TME et TILE, l'ESS mesure la perception qu'ont les sujets de leur propension à l'endormissement diurne. Validée médicalement (Johns, 1991) et conciliante vis-à-vis des algorithmes d'apprentissage automatique (cf figure 8.9), ce questionnaire semble également se conformer à l'objectif de la mesure subjective de la propension à l'endormissement pathologique. Cependant, un point bloquant dans son utilisation dans un corpus concerne les différents seuils utilisés suivant les populations, rendant confuse sa pertinence sur des populations mixtes, souvent rencontrées dans les corpus de neuropsychiatrie.

9.5.2 Classification binaire vs régression

Une façon pertinente pour les ingénieurs en apprentissage automatique de décider s'il faut binariser le problème ou le garder sous la forme d'une régression pourrait être d'imiter les cliniciens. En effet, quand certaines mesures sont reconnues comme étant catégoriques, la binarisation du problème en utilisant une valeur seuil semble pertinente. Au contraire, quand les cliniciens utilisent ces scores de manière continue, la binarisation des problèmes concernés est dénuée de sens, même si cela facilite le problème de classification par les algorithmes de classification automatique. Dans tous les cas, nous recommandons aux concepteurs de corpus d'inclure le score brut dans ceux-ci, pour permettre aux ingénieurs et aux médecins de tester différentes configurations.

9.5.3 Métadonnées

Comme mentionné par Qian *et coll.* (2020), comprendre entièrement les phénomènes mis en jeu lors de l'utilisation de ces corpus nécessite la collecte de nombreuses métadonnées sur les locuteurs et les conditions d'enregistrement. Cela permet de découvrir des biais qui n'étaient pas nécessairement identifiés et de s'assurer que lors de l'élaboration de classifieurs,

les échantillons sont classifiés par la variable désirée et non par un biais qui lui est lié (Sturm *et coll.*, 2014; Sturm, 2016)

De plus, ces mesures permettent d'étudier la robustesse des systèmes élaborés au regard d'autres variables (sexe, données démographiques, comorbidités). Par exemple, une étude portant sur la détection de la dépression dans la voix (Pan *et coll.*, 2019) a réexaminé un système de classification en prenant en compte des facteurs démographiques et l'état émotionnel des locuteurs, qui sont des facteurs confondants de la dépression. Ils ont ainsi pu parvenir à la conclusion que leur système est robuste à la fois aux données démographiques, mais aussi à l'état émotionnel des locuteurs : ce résultat n'aurait jamais pu être obtenu sans la collecte de ces informations dans ce corpus.

Nous encourageons donc l'annotation systématique des enregistrements avec un maximum d'informations pertinentes (mesures du physique du participant, comorbidités, données démographiques ...) pour à la fois évaluer la robustesse des systèmes étudiés vis-à-vis de ces données et étudier et corriger le label utilisé pour annoter la voix et en tirer des conclusions sur la manifestation des phénomènes dans la voix.

9.5.4 Conclusion et recommandations sur l'annotation des données

- **Choix de l'annotation pertinente**
  - Nous encourageons les collaborations interdisciplinaires entre les différents champs de recherche impliqués dans le choix de l'annotation (principalement ingénieur en traitement de données et médecins), permettant d'éviter un mauvais choix de label et garantissant la qualité de celui-ci (Littmann *et coll.*, 2020).
- **Fournir les scores de manière brute**
- **Collecter le maximum de métadonnées**
  - Nous encourageons la collecte d'un maximum de métadonnées sur les locuteurs ;
  - De même, systématiquement décrire la population étudiée et les conditions expérimentales dans lesquelles les locuteurs ont été enregistrés permettrait d'améliorer les comparaisons entre systèmes et l'identification des phénomènes en jeu.

Chapitre 10

L'oreille humaine est-elle capable d'estimer la somnolence dans la voix ? L'étude Endymion

Sommaire

Ce chapitre a bénéficié des travaux d'Aymeric Ferron, accueilli dans le cadre de son stage de deuxième année d'école d'ingénieur à l'École Nationale Supérieure d'Électronique, Informatique, Télécommunications, Mathématique et Mécanique de Bordeaux (ENSEIRB).

10.1 Contexte

Après avoir décrit les corpus existants et avoir discuté la méthodologie à mettre en place lors de la conception d'un corpus pour la détection de la somnolence dans la voix, ce chapitre étudie la faisabilité de la tâche de détection de la somnolence par l'oreille humaine à partir d'échantillons de la base TILE, avec un objectif double : d'une part, valider la faisabilité de la tâche de détection ; d'autre part, valider la base TILE pour une telle tâche.

10.1.1 Précédents travaux

L'estimation automatique de la somnolence subjective instantanée a fait l'objet de deux compétitions internationales en 2011 et 2019 (Schuller *et coll.*, 2011, 2019). Lors de la première compétition, la tâche consistait à classer automatiquement des échantillons de parole entre «somnolent» et «non somnolent», défini par un seuil de 7.5 sur la KSS composite proposée dans le SLC. Au cours de ce défi, le meilleur système a atteint 71.7% de précision de rappel non pondérée (UAR) (Huang *et coll.*, 2011). La deuxième compétition, en 2019, portait sur une tâche de régression visant à estimer le même score composite de somnolence que dans le premier, mais sur la base d'un nouveau corpus – le corpus SLEEP. Le meilleur système du défi a obtenu un ρ de Spearman de 38.3% (Gosztolya, 2019). Les détails de ces deux compétitions sont fournis dans le chapitre 11.

Alors qu'il existe une riche littérature sur la perception de la voix pathologique (Kreiman *et coll.*, 2007; Kreiman et Gerratt, 2000, 2005) ou de la voix des patients atteints de Parkinson (Jaywant et Pell, 2010; Sussman et Tjaden, 2012), le travail proposé par Huckvale *et coll.* (2020) est, à notre connaissance, la seule expérience perceptive disponible sur la somnolence. Constatant la différence entre les performances obtenues dans la détection automatique de la somnolence et celles obtenues dans la détection d'autres pathologies (cf chapitre 2), l'équipe de Huckvale a étudié la faisabilité de la tâche : est-il seulement possible d'estimer la somnolence par la voix ? Pour ce faire, en plus d'une analyse minutieuse du corpus SLEEP, une expérience perceptive a été menée sur 90 échantillons de même corpus qui ont été annotés par 26 auditeurs. Cette étude a conclu que l'ouïe humaine peut estimer la somnolence à partir d'échantillons de voix, et que le plafond de verre des performances d'estimation automatique repose sur le contenu du corpus, qui contient une très grande représentation de peu de locuteurs.

Cependant, ce travail souffre de trois limitations majeures. Tout d'abord, dans le corpus SLEEP, les échantillons sont courts (entre 3 et 5 secondes), alors que la durée idéale pour estimer la somnolence avec des méthodes computationnelles semble être d'environ 20 secondes (cf chapitre 9). Deuxièmement, les auteurs n'ont pas collecté d'informations sur les auditeurs, et aucune information sur les locuteurs n'est disponible sur le corpus SLEEP. Ces informations auraient pu permettre d'identifier les différents facteurs influençant les performances des auditeurs ou les caractéristiques des locuteurs les rendant correctement ou incorrectement annotés. Enfin, la troisième limite de cette étude repose sur la vérité terrain utilisée dans le corpus SLEEP, qui ne mesure pas à proprement parler la somnolence subjective.

En effet, la vérité terrain utilisée dans le corpus SLEEP est la moyenne de trois KSS, dont une seule a été remplie par les patients eux-mêmes, mesurant correctement la somnolence

subjective. Les deux autres KSS ont été remplies par des auditeurs externes, n'ayant accès qu'à une manifestation comportementale de la somnolence : ces deux KSS ont été notées – en partie – sur la base du comportement vocal du locuteur. Par conséquent, la vérité terrain du corpus SLEEP repose déjà en partie sur ce qui est exprimé par la voix, ce qui, à notre avis, facilite son estimation à l'aide d'échantillons de voix.

10.1.2 Objectif de cette étude

Cette étude vise à répondre aux trois questions suivantes :

1. Est-il possible pour l'audition humaine d'estimer la somnolence à travers des échantillons de voix ?
2. Existe-t-il des caractéristiques des auditeurs qui pourraient interagir avec leurs annotations ?
3. Existe-t-il des caractéristiques des patients qui pourraient interférer avec l'expression de la somnolence leur voix ?

En nous appuyant de la base TILE-93 (cf chapitre 7 et annexe E), qui contient les enregistrements de 93 patients hypersomniaques annotés avec des mesures objectives et subjectives de la somnolence, nous proposons dans cet article les résultats de l'étude perceptive française Endymion [1].

10.2 Méthode

10.2.1 Description des annotateurs

Soixante et onze annotateurs ont été recrutés entre le 21 juillet 2021 et le 3 septembre 2021, au LaBRI. Tous les annotateurs sont des locuteurs natifs du français. Avant l'expérience, tous les annotateurs ont déclaré qu'ils ne souffraient d'aucune déficience auditive susceptible d'interférer avec leur capacité à annoter des fichiers audio contenant de la parole. Les informations suivantes sont collectées, sur la base d'études précédentes ayant montré qu'elles sont liées à la perception de la parole :

— le sexe (Sato, 2020; Yoho *et coll.*, 2019) : « Homme », « Femme » ou « Préfère ne pas répondre » ;
— l'âge (Goy *et coll.*, 2016; Helfer et Freyman, 2014; Tremblay *et coll.*, 2021) : tranches de 5 ans de $<$ 20 ans à $\geq$90 ans ;
— dispositif d'écoute (Cooke et García Lecumberri, 2021) : casque, écouteurs ou enceintes de bonne qualité.
— sensibilité musicale (Asaridou et McQueen, 2013; Hart, 1981; Thompson *et coll.*, 2004) : « Je travaille dans le domaine de la musique ou j'ai un hobby musical » ou « Je n'ai pas de sensibilité musicale particulière ».

Ces informations ont été renseignées pendant le test sous la conduite de l'investigateur supervisant l'annotation (pas d'effet de l'investigateur dans les analyses suivantes, tests de Mann-Whitney bilatéraux, n.s.), avant les étapes de définitions et d'annotation (cf. section 10.2.3 de ce même chapitre). Ni ces informations ni l'annotation des fichiers audio n'ont été jugées comme des facteurs identifiants par les juristes rattachés à l'université de Bordeaux. Cette étude se déroule donc en dehors du cadre du régime général de protection des données

1. Moins connu que Morphée, Endymion est une figure de la mythologie grecque personnifiant le sommeil

(RGPD). Néanmoins, les volontaires ont été informés par une notice [2] du traitement ultérieur des données collectées. Le tableau 10.1 résume les données des auditeurs que nous avons collectées.

N	Âge					Sexe		Dispositif d'écoute			Sen. musicale	
	Min.	Max.	20-25	25-30	Autres	F	M	Enc.	Éc.	Cas.	Sen.	Pas de sen.
71	20-25	70-75	30	26	15	22	49	12	12	47	30	41

Tableau 10.1 – Informations collectées sur les auditeurs. *Enc.* : Enceintes ; *Éc.* : Écouteurs ; *Cas.* : Casque audio ; *Sen.* : Sensibilité.

10.2.2 Description des locuteurs et des enregistrements

Corpus TILE et métadonnées

Les enregistrements audio annotés sont extraits du corpus TILE. Dans cette étude, nous considérons les données suivantes sur les patients :

- des mesures de somnolence à court terme :
  - leurs latences d'endormissement au TILE ;
  - leurs scores à la KSS remplis au cours du TILE ;
  - leurs scores sur l'échelle des visages [CFS, (Maldonado *et coll.*, 2004)] ;
- des facteurs confondants à la fois de la somnolence et la voix :
  - leur score à l'échelle de sévérité de la fatigue [FSS, (Krupp *et coll.*, 1989)], mesurant l'impact de la fatigue sur le fonctionnement quotidien ;
  - leur score à l'échelle d'activation de l'hôpital de Toronto [THAT, (Shahid *et coll.*, 2016)], estimant le niveau d'activation habituel des patients ;
  - les sous-échelles respectives de l'échelle hospitalière d'anxiété et de dépression [HAD-A et HAD-D, (Zigmond et Snaith, 1983)] ;
- des mesures de somnolence excessive :
  - l'échelle de somnolence d'Epworth [ESS, (Johns, 1991)], un questionnaire en 8 items mesurant la propension à l'endormissement diurne dans des conditions inadéquates ;
  - l'index de somnolence de Barcelone [BSI, (Guaita *et coll.*, 2015)], un questionnaire en deux items mesurant la SDE ;
- des mesures physiques et sociodémographiques :
  - age, sexe, IMC, circonférence du cou ;
  - niveau sociodémographique (nombre d'années d'études après le brevet) ;

Sous-corpus Endymion

Afin de réduire la variabilité des annotations dues au grand nombre de locuteurs dans le corpus TILE, nous avons regroupé 20 patients (10 hommes et 10 femmes) dans un sous-corpus, appelé sous-corpus Endymion dans la suite.

Critères de sélection Les patients du sous-corpus ont été sélectionnés de manière à ce que leurs variations de somnolence objective (latence d'endormissement) et subjective (KSS et CFS) au cours des siestes soient maximales. Ainsi, lors de l'annotation de tous les échantillons d'un

2. Disponible en ligne

locuteur dans le cadre du paradigme Baseline (cf section 10.2.3), les annotateurs sont exposés à la plus grande variété possible d'états pour un même patient, ce qui facilite leur différenciation des différents états du locuteur et donc leur annotation. En effet, si un locuteur avec un niveau de somnolence constant avait été proposé pour l'annotation, l'auditeur aurait pu se tromper, s'attendant à être exposé à différents niveaux de somnolence. De plus, tous les locuteurs sélectionnés ont au moins une sieste pendant laquelle ils ne se sont pas endormis, afin que celle-ci soit utilisée comme référence pendant le paradigme Baseline. Pour les mêmes raisons, deux locuteurs supplémentaires (un homme et une femme) ayant les variations les plus élevées sur le TILE, KSS, et CFS ont été particulièrement sélectionnés pour être les locuteurs d'entraînement dans le paradigme Baseline. Ces deux locuteurs ne sont annotés dans aucune phase de test et sont toujours les mêmes dans la phase d'entraînement du paradigme Baseline. Leurs enregistrements ne sont pas utilisés dans le paradigme Random. Le tableau 10.2 résume les caractéristiques des locuteurs du sous-corpus Endymion utilisé dans cette étude.

	Sexe F/M	Âge	IMC (kg/m^2)	Cou (cm)	Édu. (ans)	FSS (9-63)	THAT (0-50)	HADA (0-21)	HADD (0-21)	BSI (0-6)	ESS (0-24)	TILE (0-20 min)
base TILE ($n = 93$)	58/35	36.6 (14.4)	23.9 (5.1)	37.8 (4.4)	5.5 (2.6)	48.8 (10.5)	22.8 (7.2)	8.5 (4.2)	6.7 (3.8)	2.3 (1.0)	14.6 (4.7)	11.6 (4.6)
Endymion ($n = 20$)	10/10	42.6 (12.8)	25.3 (6.0)	38.9 (4.7)	5.6 (2.8)	49.2 (11.6)	20.8 (6.1)	7.5 (3.5)	7.0 (4.5)	2.2 (1.0)	14.6 (4.5)	14.3 (2.9)

TABLEAU 10.2 – Caractéristiques des locuteurs du corpus TILE et du sous-corpus Endymion.
Édu. : niveau d'éducation ; *Cou* : Circonférence du cou. Valeurs données sous la forme moy. (é-t).

Corrélation entre les caractéristiques des locuteurs Certains des facteurs décrits ci-dessus peuvent être interdépendants en raison du processus d'échantillonnage impliqué lors de la collecte du corpus, ou lors de la sélection du sous-corpus. Pour mettre en lumière ces corrélations, nous avons calculé, pour chaque paire de facteurs, leur corrélation (ρ de Spearman). Ainsi, dans le sous-corpus Endymion, la circonférence du cou des patients est significativement corrélée à l'âge ($\rho = .47, p = .03$) et à l'IMC ($\rho = .66, p < 10^{-3}$), ce qui est conforme à ce qui est observé sur la population générale (Ben-Noun *et coll.*, 2001). Les scores de dépression et d'anxiété sont corrélés ($\rho = .63, p = 4.0 \times 10^{-3}$), ainsi que les scores du BSI et de l'ESS ($\rho = .56, p = 2.0 \times 10^{-3}$), ce qui est une conséquence directe de la conception de ces questionnaires (Guaita *et coll.*, 2015; Zigmond et Snaith, 1983).

10.2.3 Description de l'expérience

Prétraitement des échantillons audio

Dans le corpus TILE, les enregistrements sont longs (durée moyenne : 77 secondes). Avec des tels échantillons, il serait difficile voire impossible pour l'annotateur de rester pleinement concentré non seulement pendant toute la durée de l'échantillon, mais pendant toute la session d'annotation, qui durerait alors plus d'une heure. Afin de prendre en compte ce facteur, nous avons découpé les enregistrements et conservé les 30 premières secondes de chaque enregistrement (moy. : 32.0 s, é-t : 1.3 s). Afin de conserver une fluidité de lecture naturelle pour les annotateurs et d'éviter de couper l'enregistrement au milieu d'une phrase, nous avons utilisé le système d'alignement de pauses de lecture décrit dans le chapitre 15. Enfin, les échantillons audio sont tous normalisés à -3 dB afin d'assurer un volume sonore constant entre les échantillons.

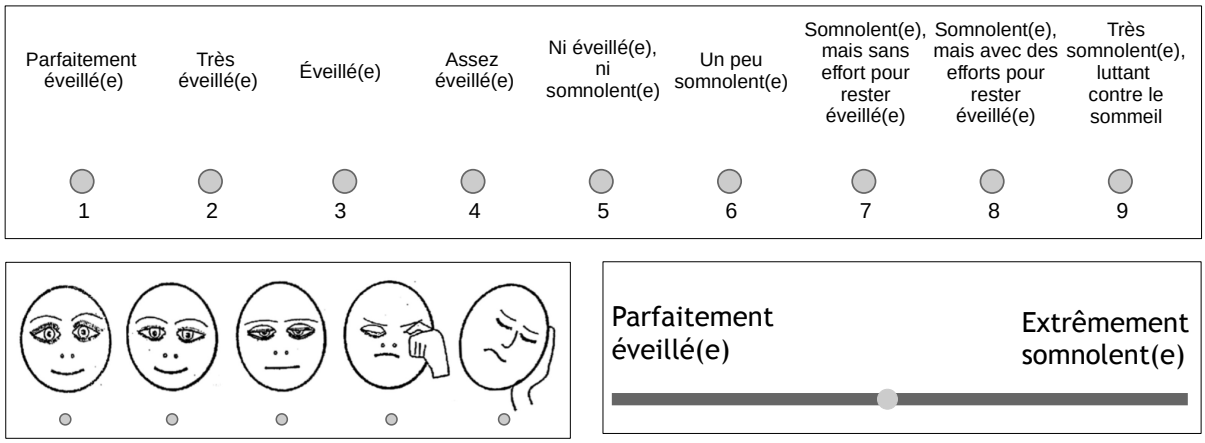

FIGURE 10.1 – Outils d'annotation proposés aux annotateurs de l'étude. Haut : version française de la KSS en version inclusive. Gauche : CFS. Droite : Curseur.

Paradigmes : Random vs. Baseline

Deux paradigmes d'annotation sont considérés dans cette étude. Dans le paradigme Random, les fichiers à annoter sont choisis de manière pseudo-aléatoire dans le sous-corpus Endymion, en favorisant les fichiers les moins annotés. Ainsi, chaque fichier est complètement indépendant du précédent, et les annotations de l'auditeur sont décorrélées de l'ordre des échantillons. Dans ce paradigme, les annotateurs traitent 5 fichiers pendant la phase d'entraînement, et 10 fichiers différents pendant la phase de test.

Pendant le paradigme Baseline, les patients annotent 4 échantillons du même locuteur, avec l'aide d'un enregistrement de référence du même locuteur lorsqu'il est éveillé (la *baseline*). Cette session a été choisie comme étant la session la plus précoce pendant laquelle le patient ne s'est pas endormi (latence d'endormissement au TILE de 20 minutes). Dans ce paradigme, les auditeurs annotent les fichiers des deux patients précédemment choisis pour la phase d'entraînement, et de deux autres patients choisis au hasard parmi les moins annotés. Pour chaque locuteur, l'ordre des échantillons à annoter est choisi aléatoirement afin de rompre toute corrélation liée à l'ordre des fichiers audios.

Outils d'annotation et tâches

Trois outils d'annotation avec trois instructions différentes correspondant à trois tâches différentes ont été choisis pour chaque scénario. Ils sont représentés dans la figure 10.1.

Somnolence subjective – KSS-a Une des tâches d'annotation proposées aux auditeurs est d'essayer de retrouver, à partir des enregistrements vocaux, le score KSS que le patient s'est attribué immédiatement après la lecture. Contrairement au paradigme proposé dans (Huckvale *et coll.*, 2020), il ne s'agit pas d'estimer les manifestations comportementales de la somnolence du locuteur à l'aide d'une KSS, mais d'estimer sa perception de son propre niveau de somnolence. La KSS utilisée est la traduction française réalisée par l'hôpital de Bordeaux, avec une description textuelle de tous les niveaux. Elle est proposée dans sa version inclusive [3]. Dans la

3. La version officielle de la KSS en Français n'existe pas dans les bases de la *Mappi research trust*. Cependant, puisque les versions officielles de l'ESS et de l'ISI sont proposées en version inclusives, nous avons pris le parti de proposer aux annotateurs la KSS sous cette même forme.

suite, le score à la KSS annotée par l'auditeur est appelé KSS-a (KSS-annotation) pour le différencier de la valeur de vérité terrain annoté par les patients. Un tableau de correspondance entre les deux versions (anglaise et française) de ce questionnaire est proposé en Annexe C.

Somnolence subjective – CFS-a Une deuxième tâche proposée consiste à estimer dans la voix du patient son autoannotation sur la CFS. Comme pour la tâche précédente, le but n'est pas d'estimer la somnolence du locuteur, mais sa perception de son niveau de somnolence. Pour ce faire, les annotateurs utilisent l'échelle des visages, avec le visage le plus éveillé à gauche et le plus endormi à droite. Sans texte, cette échelle a été conçue pour permettre une évaluation plus intuitive de la somnolence. Pour des raisons pratiques, les visages annotés sont ensuite mis en correspondance avec une échelle de 0 à 4. Comme pour la KSS, le score CFS annoté par l'auditeur est appelé CFS-a (CFS-annotation) dans la suite.

Latence d'endormissement – Curseur La dernière tâche évaluée est l'estimation de la latence d'endormissement des patients, en demandant aux annotateurs d'évaluer « la propension à l'endormissement du patient. » À cette fin, l'outil proposé est un curseur à 100 niveaux dont les extrémités sont étiquetées avec « Parfaitement éveillé » à gauche et « Extrêmement somnolent » à droite.

Scénario de l'expérience

La chronologie de chaque test est la même pour chaque annotateur et est décrite ci-dessous.

1. Notice d'information, consentement et réglage du niveau sonore : l'auditeur lit la notice d'information sur l'expérience, accepte de participer en cochant une case. Il peut écouter librement un clip audio de test, également normalisé à -3dB, pour ajuster le niveau sonore de son dispositif d'écoute, dans le but d'entendre clairement et distinctement la voix contenue dans l'enregistrement.
2. Remplissage du formulaire : l'annotateur remplit le formulaire recueillant les informations le concernant.
3. Présentation des définitions et de l'outil d'annotation : l'annotateur est invité à lire un écran contenant la définition de la somnolence employée dans cette étude (à savoir la propension à l'endormissement) et la différence entre somnolence et fatigue :

   > « Dans cette étude, *la somnolence* sera définie comme la tendance d'une personne à s'endormir, à ne pouvoir résister contre le sommeil. La remédiation à un état de somnolence est le sommeil. Nous insistons sur l'opposition entre *somnolence* et *fatigue*. Par opposition à la somnolence, la remédiation à un état de fatigue ou d'épuisement est le repos. »

   Enfin, une brève présentation de la tâche et de l'outil d'annotation est affichée. À la fin de cet écran, l'annotateur peut librement essayer l'outil d'annotation. Durant toutes les phases précédentes, l'annotateur est libre de poser toute question. À la fin de cette partie, il est informé que l'interaction avec l'investigateur ne sera pas autorisée dans les étapes suivantes, sauf en cas de problème logiciel.
4. Phase d'annotation : l'annotation se fait en deux étapes.
   (a) Une étape d'entraînement, au cours de laquelle les échantillons sont présentés un par un. Après chaque annotation, la vérité terrain et l'annotation proposée sont

données côte à côte. Pendant cette étape, l'annotateur traite soit 10 échantillons tirés aléatoirement de la base de données si l'expérience se déroule dans le cadre du paradigme Random, soit les enregistrements des deux locuteurs dédiés pour le paradigme Baseline ;

(b) Une étape de test, au cours de laquelle l'auditeur estime le niveau de somnolence soit de 10 enregistrements aléatoires (paradigme Random), soit des fichiers audio correspondant à deux locuteurs (paradigme Baseline). Contrairement à la phase d'entraînement, les valeurs de vérité terrain ne sont pas données à l'annotateur au fur et à mesure du processus d'annotation.

Deux scénarios par annotateur

Afin de profiter au maximum de la présence des annotateurs tout en proposant des sessions d'annotation de durées raisonnables, nous avons demandé à chaque annotateur d'annoter deux scénarios différents, l'un dans le paradigme Random, l'autre dans le paradigme Baseline. Afin de prendre en compte l'influence de l'ordre, la moitié des annotateurs ($n = 35$) a d'abord annoté le paradigme Baseline puis le paradigme Random, tandis que l'autre moitié ($n = 36$) a fait l'inverse. De plus, pour différencier les deux scénario autant que possible, les outils d'annotation étaient différents dans les deux scénarios : « Random/curseur » était apparié avec « Baseline/CFS », « Random/CFS » avec « Baseline/KSS », et « Random/KSS » avec « Baseline/curseur ». Aucun effet significatif de l'ordre n'a été observé dans les résultats suivants (test de Mann-Whitney bilatéral, n.s.).

10.2.4 Méthodes d'analyse des données

Normalisation

Les annotations effectuées par les auditeurs passent par trois normalisations différentes, afin d'obtenir des scores de pourcentage de somnolence entre 0 et 1 qui soient à la fois indépendants des annotateurs, indépendants de la plage de somnolence à laquelle ils ont été exposés, et enfin indépendants de l'outil d'annotation utilisé. Ces trois étapes sont représentées dans la figure 10.2.

1. La première normalisation est une *z-normalisation* (centrage de la moyenne à 0 et réduction de l'écart-type à 1) appliquée aux annotations de chaque auditeur afin de supprimer son comportement d'annotation spécifique.
2. Ensuite, pour tenir compte de la disparité d'amplitude des états de somnolence présentés aux auditeurs, nous normalisons les scores précédents afin que le minimum et le maximum correspondent à leurs valeurs respectives dans la vérité terrain des échantillons annotés par l'auditeur (normalisation des limites).
3. Enfin, pour assurer une comparaison équitable entre les différents outils d'annotation qui ont des valeurs d'amplitudes différentes, nous normalisons l'ensemble des scores avec les valeurs minimales et maximales de l'outil utilisé (respectivement 1 et 9 pour la KSS, 0 et 100 pour le curseur, et 0 et 4 pour la CFS). La même normalisation est appliquée aux scores de la vérité du terrain. Pour être comparés à d'autres échelles, le score annoté sur le curseur et la latence d'endormissement de la vérité terrain correspondante sont inversés, de sorte que 0 correspond à une latence de sommeil de 20 minutes (le patient ne s'est pas endormi) et 1 correspond à une latence de sommeil de 0 minute (les patients se sont endormis immédiatement).

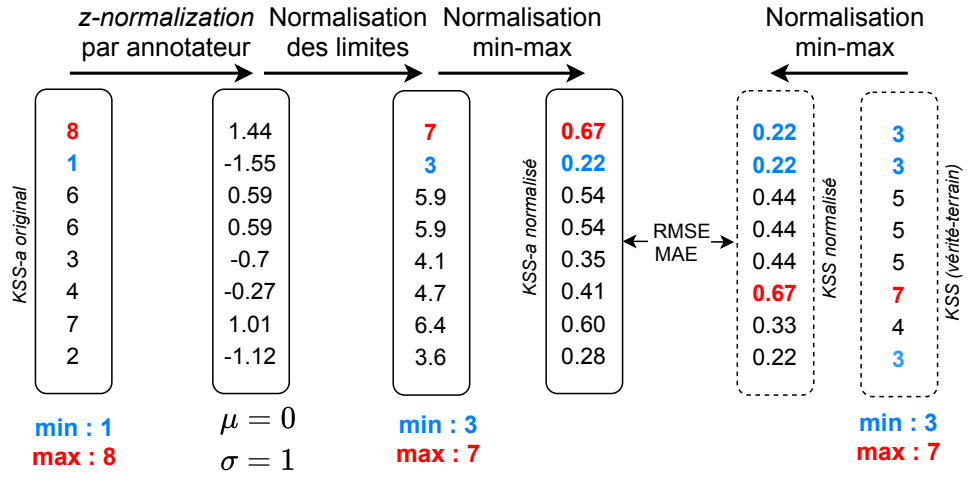

FIGURE 10.2 – Trois étapes de normalisation des annotations : *z-normalisation* par locuteur, normalisation des limites et normalisation min-max.

Le score résultant est compris entre 0 et 1 et représente un «pourcentage de somnolence», qui peut être comparé entre les échelles d'annotation.

Mesures de la performance d'annotation

Pour comparer les performances entre les annotateurs, il faut définir une fonction de performance entre leurs annotations et la vérité terrain attendue.

La corrélation aurait pu être une bonne mesure de performance, mais le nombre d'échantillons annotés par chaque auditeur pendant la phase d'annotation est trop faible (10 échantillons pour le paradigme Random, 8 échantillons pour le paradigme Baseline) et la CFS ne comporte que cinq niveaux différents, ce qui le place à la limite entre une échelle continue et une variable catégorielle (Rhemtulla *et coll.*, 2012).

Par conséquent, nous préférons utiliser des mesures basées sur l'erreur absolue, définie comme la différence absolue entre la valeur de la vérité terrain y et l'annotation proposée $\hat{y}$. Il existe deux métriques classiques pour mesurer les performances basées sur l'erreur absolue : l'erreur absolue moyenne (MAE), qui est la moyenne arithmétique de l'erreur absolue sur l'ensemble des échantillons, et l'erreur quadratique moyenne (RMSE), définie de manière similaire, mais avec une moyenne géométrique de l'erreur absolue :

$$\text{MAE} = \frac{1}{n}\sum_k |\hat{y}_k - y_k|$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_k |\hat{y}_k - y_k|^2}$$

Par conséquent, des RMSE et MAE plus faibles signifient des erreurs plus faibles, et donc de meilleures performances d'annotation. Le choix entre les deux mesures a fait l'objet d'intenses discussions dans la littérature (Chai et Draxler, 2014; Willmott et Matsuura, 2005) sans qu'un consensus ait pu émerger : afin d'obtenir une meilleure généralisation et une plus grande robustesse dans nos résultats, nous choisissons de reporter les deux.

Exemple : En reprenant les scores proposés dans la figure 10.2, la MAE et la RMSE se calculent de la manière suivante :

$$\begin{aligned}\text{MAE} &= \frac{1}{8}(|0.67-0.22|+|0.22-0.22|+|0.54-0.44|+...+|0.22-0.28|)\\ &= \frac{1}{8}\times 1.77 = 0.22\end{aligned}$$

$$\begin{aligned}\text{RMSE} &= \sqrt{\frac{1}{8}(|0.67-0.22|^2+|0.22-0.22|^2+|0.54-0.44|^2+...+|0.22-0.28|^2)}\\ &= \sqrt{\frac{1}{8}\times 0.375} = 0.216\end{aligned}$$

10.2.5 Performances en fonction du paradigme et de l'outil d'annotation

La première partie de l'analyse concerne les performances d'annotation indépendamment de tout facteur. Afin d'étudier l'influence de l'outil d'annotation ou du paradigme sur les performances, nous calculons la RMSE et la MAE entre l'annotation des auditeurs et les valeurs attendues de la vérité terrain pour toutes les combinaisons possibles d'outils d'annotation et de paradigmes.

10.2.6 Performances des annotateurs en fonction de leurs caractéristiques

Ensuite, afin d'étudier les caractéristiques des annotateurs qui influencent leurs annotations, nous calculons la RMSE et la MAE entre les annotations et la vérité terrain pour chaque annotateur et chaque paradigme.

Expérience précédente

Chaque annotateur complète deux scénarios différents avec des paradigmes et des outils d'annotation différents. Un biais potentiel introduit par une telle pratique est une éventuelle amélioration de la performance dans le second test grâce à l'expérience acquise lors du premier. Pour tester cette hypothèse, nous calculons un test de Mann-Whitney bilatéral entre les deux distributions « RMSE - Première utilisation » et « RMSE - Deuxième utilisation » pour chaque combinaison de paradigme et d'outil d'annotation. Nous procédons de la même manière pour la MAE.

Performance des annotateurs

Pour mettre en lumière la relation entre les caractéristiques des annotateurs et leurs performances, nous calculons des tests de Mann-Whitney bilatéraux sur leur RMSE et MAE entre chaque paire de valeurs de chaque caractéristique mesurée. Cette procédure est appliquée pour chaque combinaison possible d'outils d'annotation et de paradigmes. Les caractéristiques étudiées sont celles décrites précédemment : sexe, catégorie d'âge, dispositif d'écoute, et sensibilité musicale.

Capacité des annotateurs à annoter la somnolence à partir d'échantillons de voix

Afin d'évaluer la capacité de l'oreille humaine d'estimer la somnolence à partir d'échantillons de voix, nous calculons la MAE et la RMSE dans les conditions les plus favorables en fonction des facteurs interférant avec les performances des annotateurs précédemment identifiés.

10.2.7 Caractéristiques des locuteurs et qualité des annotations

Nous étudions également les caractéristiques des locuteurs qui pourraient favoriser ou interférer avec la reconnaissance de la somnolence par les annotateurs. Pour ce faire, nous calculons, pour chaque locuteur, la RMSE et la MAE entre les annotations qui ont été proposées et les vérités terrain correspondantes. Ensuite, nous calculons la corrélation (ρ de Spearman) entre les caractéristiques des patients et leurs performances associées pour chaque combinaison possible de paradigme et d'outils d'annotation. Les facteurs considérés sont ceux décrits précédemment : sexe, âge, IMC, tour de cou, niveau d'éducation, niveau de fatigue (FSS), de vigilance (THAT), d'anxiété (HAD-A), de dépression (HAD-D), de SDE (BSI et ESS) et de propension à l'endormissement (TILE).

10.3 Résultats

10.3.1 Performances en fonction du paradigme et de l'outil d'annotation

Les RMSE et MAE entre les annotations et les valeurs attendues de la vérité terrain pour toutes les combinaisons possibles d'outils et de paradigmes d'annotation sont présentées dans le tableau 10.3.

Paradigme	Métrique	KSS-a	CFS-a	Curseur
Random	RMSE	**.27**	.32	.40
	MAE	**.30**	.31	.47
	n	220	250	240
Baseline	RMSE	**.26**	.28	.39
	MAE	.29	**.28**	.45
	n	192	192	184
Random + Baseline	RMSE	**.26**	.30	.39
	MAE	**.29**	.30	.46
	n	412	442	424

Tableau 10.3 – RMSE et MAE entre les annotations et la vérité terrain pour chaque outil d'annotation, en fonction du paradigme.

Outils d'annotation ou tâche

L'effet prédominant sur les performances est l'outil d'annotation ou la tâche : les meilleures MAE (.31-.28) et RMSE (.32-.26) sont obtenues en utilisant le KSS-a ou le CFS-a – avec peu de différences entre eux – alors que les performances du curseur sont nettement inférieures (.47-.39).

Paradigme

Le deuxième effet prédominant est une différence faible, mais notable entre le paradigme Random et le paradigme Baseline, ce dernier conduisant à de meilleures RMSE (-1% pour le KSS-a, -4% pour le CFS-a, -1% pour le curseur) et MAE (-1% pour le KSS-a, -3% pour le CFS-a, -2% pour le curseur).

10.3.2 Performances des annotateurs en fonction de leurs caractéristiques

Expérience précédente

Nous n'avons pas observé de différence significative (tests bilatéraux de Mann-Whitney, tous les $p > .05$) entre l'annotation faite pendant la première ou la deuxième session d'annotation, pour toute combinaison d'outil d'annotation et de paradigme, à la fois pour la RMSE et la MAE. Par conséquent, dans les résultats suivants, les deux sessions d'annotation de chaque annotateur sont considérées comme deux sessions indépendantes.

Performance des annotateurs

Les caractéristiques des annotateurs conduisant à des différences significatives ($p < .05$) entre les performances sur chacune de leur paire de valeurs sont rapportées dans le tableau 10.4.

	Métrique	Casque audio *moy. (é-t)*	Écouteurs *moy. (é-t)*	MW U	p
Random		$n = 92$	$n = 24$		
	RMSE	.34 (.11)	.41 (.09)	1468	**.01**
	MAE	.27 (.10)	.34 (.09)	1498	**.007**
CFS-a		$n = 58$	$n = 22$		
	RMSE	.26 (.10)	.32 (.07)	874	**.01**
	MAE	.20 (.09)	.26 (.05)	918	**.003**

TABLEAU 10.4 – Différences significatives dans les performances des annotateurs en fonction de leur dispositif d'écoute ($p < .05$). MW : test de Mann-Whitney.

La seule différence significative observée concerne le dispositif d'écoute : les auditeurs qui ont utilisé un casque ont des RMSE et des MAE plus faibles que leurs homologues qui ont utilisé des écouteurs pour le paradigme aléatoire (+7% de RMSE et de MAE) ou l'annotation avec le CFS-a (+6% de RMSE et de MAE). À l'exception du dispositif d'écoute, aucune influence d'une autre caractéristique de l'annotateur n'a pu être identifiée.

Capacité des annotateurs à annoter la somnolence à partir d'échantillons de voix

Dans une configuration prenant en compte les précédentes observations sur les effets du dispositif d'écoute, de l'outil d'annotation et du paradigme en utilisant un casque audio, annoter les échantillons avec la CFS sous le paradigme Baseline conduit à une MAE par locuteur moyen de 16.5% (é-t : 7.7%) et une RMSE moyenne équivalente de 22.9% (std : 8.0%).

10.3.3 Caractéristiques des locuteurs et qualité des annotations

Les caractéristiques des locuteurs qui sont significativement corrélées à la fois à la RMSE et à la MAE ($p < .05$) sont indiquées dans le tableau 10.5.

Facteur	Paradigme	Outils	RMSE		MAE	
			ϱ	p	ϱ	p
Fatigue	Random	CFS	.52	.02	.51	.02
		Tous	.49	.03	.48	.04
Anxiété	Random	Curseur	-.51	.02	-.56	.01
		Tous	.46	.04	.49	.03
Édu.	Random	Curseur	-.46	.04	-.47	.04
	Baseline		-.53	.02	-.57	.009
	Tous		-.54	.01	-.57	.009

Tableau 10.5 – Corrélations significatives (Spearman ϱ, $p < .05$) entre les caractéristiques des locuteurs et leurs RMSE et MAE associées.
Par. : Paradigme ; *Outils* : Outils d'annotation ; *Édu.* : Niveau d'éducation.

Le niveau de fatigue des locuteurs influe sur les performances d'annotation dans le paradigme Random : lorsque les locuteurs rapportent des niveaux de fatigue élevés (score à la FSS élevé), les annotateurs qui annotent leurs échantillons font plus d'erreurs (c'est-à-dire que la RMSE et la MAE sont plus élevées). Cette différence n'a pas été observée dans le paradigme Baseline. De même, le niveau d'anxiété du locuteur (mesuré par le score à la HAD-A) interfère avec la performance des annotateurs dans le paradigme Random. Enfin, le niveau d'éducation du locuteur est anticorrélé avec la RMSE et la MAE sur tous les paradigmes utilisant le curseur : lorsque le niveau d'éducation du locuteur augmente, les erreurs commises par les auditeurs lors de l'estimation de sa latence de sommeil à l'aide du curseur diminuent.

10.4 Discussion

10.4.1 Performances en fonction du paradigme et de l'outil d'annotation

Outils d'annotation ou tâche

L'estimation de la latence de sommeil des locuteurs avec un curseur a conduit à une annotation moins exacte que l'estimation de la somnolence subjective des locuteurs avec une KSS ou une CFS. Nous proposons ici deux hypothèses.

Premièrement, le curseur a été signalé par les annotateurs comme étant « plus difficile à utiliser » que la KSS ou la CFS en raison de sa granularité élevée. La CFS, la KSS et le curseur peuvent être assimilés à des échelles de type Likert avec respectivement 5, 9 et 100 niveaux différents. À cet égard, elles partagent les mêmes débats sur le nombre de niveaux que les échelles de Likert : si le fait de proposer plus de niveaux donne aux annotateurs « plus de variétés d'options qui, à leur tour, augmentent la probabilité de répondre à la réalité objective des gens » (Joshi *et coll.*, 2015), la validité en test-retest et la cohérence inter-items de ces échelles diminuent avec le nombre de niveaux proposés (Preston et Colman, 2000). À cet égard, les performances inférieures obtenues avec le curseur pourraient être dues à l'outil d'annotation, qui ne permet pas aux annotateurs de rapporter efficacement leur jugement sur le fichier audio.

La deuxième hypothèse est plus spécifique à notre tâche. En effet, nous abordons le problème de l'estimation de la somnolence à partir d'échantillons vocaux en utilisant deux opérationnalisations différentes de la somnolence, à savoir l'évaluation subjective par les locuteurs eux-mêmes et leur latence d'endormissement après la lecture d'un texte, évaluée par des mesures polysomnographiques. Nous supposons que le phénomène qui se manifeste de manière prédominante à travers la voix est le ressenti subjectif des patients sur leur niveau de somnolence au détriment de leur niveau physiologique de somnolence. Selon cette hypothèse, la tâche d'estimation de la latence de sommeil d'un patient après sa lecture en se basant uniquement sur des échantillons de voix est une tâche beaucoup plus difficile que l'estimation de l'autoannotation de leur somnolence subjective par les patients, ce qui explique les moins bonnes performances dans cette tâche.

Paradigme

De la même manière, nous proposons deux hypothèses pour expliquer l'amélioration des performances entre le paradigme Random et le paradigme Baseline.

Premièrement, nous faisons l'hypothèse que ces différences sont dues à la présence du fichier audio de référence contenant la voix du patient lorsqu'il ne s'endort pas après la lecture du texte. En effet, en choisissant un échantillon audio aléatoire sans référence, il est presque impossible de distinguer l'état cible du locuteur – dans notre cas, l'état de somnolence – de tous les autres traits exprimés par la voix, tels que l'anxiété (Baird *et coll.*, 2020), la dépression (Cummins *et coll.*, 2015), le niveau de lecture (cf. chapitre 13), etc. En proposant une telle référence, l'annotateur peut estimer la somnolence du patient indépendamment des autres traits qui pourraient s'exprimer dans la voix de ce dernier, augmentant ainsi les performances d'annotations.

Une autre hypothèse repose sur le fait d'annoter quatre fois de suite le même locuteur dans quatre états différents, qui pourrait renforcer cette distinction entre l'expression de l'état et des traits du locuteur par la voix.

10.4.2 Performances des annotateurs en fonction de leurs caractéristiques

10.4.3 Influence du sexe, de l'âge et de la sensibilité musicale

Contrairement à la littérature, qui montre un impact de ces facteurs sur le traitement de la parole, nous n'avons pas trouvé d'influence du sexe, de l'âge ou de la sensibilité musicale sur les performances. Ce résultat assure cependant que les annotations recueillies dans notre étude - et par conséquent l'étude des caractéristiques du locuteur - sont indépendantes de ces traits de l'auditeur.

Capacité des annotateurs à annoter la somnolence à partir d'échantillons de voix

Les performances obtenues dans la configuration la plus favorable (c'est-à-dire avec un annotateur utilisant un casque et annotant avec la CFS sous le paradigme Baseline) atteignent une MAE inférieure à 20% sur la CFS, c'est-à-dire inférieure à un item sur l'échelle. En conséquence, lorsqu'elle est évaluée dans des conditions minimisant les facteurs interférant avec la qualité d'annotation, l'estimation de la somnolence à partir d'échantillons vocaux semble possible.

10.4.4 Caractéristiques des locuteurs et qualité des annotations

Nos résultats suggèrent que la fatigue et l'anxiété interfèrent avec l'annotation correcte de la somnolence à partir d'échantillons vocaux dans le paradigme Random, mais pas dans le paradigme Baseline : ce dernier permet effectivement aux auditeurs d'annoter les états de somnolence des locuteurs indépendamment de leurs facteurs traits tels que la fatigue ou l'anxiété.

Une exception est cependant observée avec le niveau d'éducation des locuteurs, qui interfère avec l'annotation de la latence d'endormissement (en utilisant le curseur) sur tous les paradigmes (Random, Baseline et les deux). Nous faisons l'hypothèse que dans cette tâche, la somnolence objective a le même impact sur la voix qu'un niveau d'éducation plus faible : les patients somnolents font plus d'erreurs de lecture et ont des comportements de lecture différents concernant la localisation et la durée des pauses de lecture (cf chapitre 15).

10.5 Limites et perspectives

Bien que nos résultats suggèrent que l'oreille humaine est capable de distinguer la somnolence à partir d'échantillons vocaux, cette étude n'est que la deuxième à aborder le lien entre la somnolence et la perception de la voix. Par conséquent, nous proposons le programme de recherche suivant afin de combler les lacunes concernant l'estimation humaine de la somnolence par la voix.

10.5.1 Dissociation entre la tâche et l'outil d'annotation

Comme il y avait une corrélation entre la tâche et les outils d'annotation dans cette expérience, nous n'avons pas pu expliquer l'origine des performances plus faibles de l'estimation de la latence d'endormissement avec un curseur. Une nouvelle étude décorrélant la tâche d'annotation et l'outil proposé aux auditeurs – par exemple, en utilisant uniquement un curseur pour estimer soit la somnolence subjective des patients, soit la latence d'endormissement, ou en utilisant les différents outils d'annotation pour annoter une seule tâche – devrait permettre de savoir qui de la tâche ou de l'outil d'annotation a un impact plus important sur la performance.

10.5.2 Influence dans la langue parlée sur les performances

Alors que la présente étude a été menée avec des auditeurs français annotant des échantillons de voix enregistrées par des locuteurs français, l'étude menée par Huckvale *et coll.* (2020) a demandé à des auditeurs britanniques d'annoter des échantillons enregistrés par des locuteurs allemands. Le niveau de compréhension de la langue peut permettre – ou empêcher – l'annotateur d'utiliser des indices linguistiques pour estimer le niveau de somnolence des locuteurs et devrait être étudié plus avant. Des études croisant les langues des auditeurs et des locuteurs, ou mesurant la maîtrise de la langue parlée par l'auditeur, pourraient apporter des résultats intéressants sur la façon dont nous estimons la somnolence à partir du discours.

10.5.3 Durée des échantillons

Dans la présente étude, nous nous sommes concentrés sur des échantillons d'au moins 30 secondes, alors que le corpus de référence SLEEP comprend des échantillons qui sont tous inférieurs à 5 secondes. Un plan expérimental spécifique devrait permettre de dégager un

consensus sur la longueur minimale d'enregistrement obligatoire pour estimer la somnolence avec l'audition humaine.

10.5.4 Profil des annotateurs

Enfin, une comparaison entre différentes populations d'auditeurs habitués à entendre des voix somnolentes devrait apporter des informations intéressantes sur les processus sous-jacents de l'annotation de la somnolence, et sur la façon dont l'audition humaine estime la somnolence. Par exemple, la comparaison des annotations faites par des orthophonistes – qui ont une oreille très entraînée et une bonne connaissance de leurs propres processus d'écoute, mais ne sont pas spécialisés dans la somnolence – et des médecins du sommeil – qui ont l'habitude de mener des entretiens cliniques avec des patients somnolents, mais ne se concentrent généralement pas sur ce qu'ils entendent – devrait donner des résultats intéressants.

Conclusion de la partie

Comparaison des bases de données de l'état de l'art et recommandation de construction de corpus

Dans cette partie nous avons décrit précisément les quatre corpus les plus utilisés de l'état de l'art sur la détection de la somnolence dans la voix, avec une attention particulière donnée au corpus TILE, qui sera exploité dans les parties suivantes de ce manuscrit.

En les comparant entre eux, nous avons pu identifier deux grandes catégories de corpus :

1. Les corpus ayant pour but la détection dans la voix de la somnolence subjective au court terme de sujets placés en privation de sommeil, pour lesquels la somnolence est induite par cette privation (SLC et SLEEP) ;
2. Les corpus ayant pour but la détection dans la voix de la somnolence objective au long terme de patients hypersomniaques passant un test médical, pour lesquels la somnolence est induite par une pathologie (TME et TILE).

À partir de la précédente comparaison, nous avons établi des recommandations sur la construction de tels corpus, à la fois sur la conception de l'expérience globale (nombre d'enregistrements par sujet, nombre de sujets, population incluse, lieu(x) d'enregistrement, qualité d'enregistrement), sur la conception de la tâche vocale (tâche proposée, durée) mais aussi sur l'annotation des données avec une vérité terrain, qui est soumise à trois contraintes : validité médicale, équilibre des classes et adéquation avec l'objectif.

Étude perceptuelle Endymion

Par ailleurs, l'étude perceptuelle Endymion a permis de valider la faisabilité de la tâche de détection de la somnolence subjective à partir d'échantillons vocaux par l'oreille humaine sur la base TILE, ouvrant la voie à sa détection par des algorithmes automatiques.

En revanche, la tâche équivalente sur la somnolence objective n'a pas été concluante : les annotateurs naïfs inclus dans l'étude n'ont pas été capables d'estimer la latence d'endormissement au TILE à partir d'enregistrements vocaux. La faisabilité de cette tâche par des algorithmes automatiques est donc incertaine.

Base SOMVOICE

Enfin, une nouvelle base de données, suivant strictement les recommandations effectuées dans cette partie, est en cours de collecte à la plateforme de recherche en Sommeil, Addiction et Neuropsychiatrie (SANPSY), ajoutant aux critères d'inclusion et d'exclusion habituellement utilisés dans les études sur le sommeil des critères de niveau de lecture, qui ont été validés par les comités d'éthique et de méthodologie. De plus, un nouvel ensemble de textes a été sélectionné en collaboration avec les stagiaires en orthophonie encadrées au cours de cette thèse, prenant en compte les différentes lacunes des textes précédemment proposés pour la base TILE (notamment sans dialogues).

Enfin, cette étude permettra de faire le lien entre les catégories de corpus précédemment mentionnées : le paradigme central est basé sur des TILE effectués par des sujets sains avec

et sans privation de sommeil et comprend à la fois des mesures objectives (TILE et test psychomoteur de vigilance) et subjectives (KSS, Échelle Visuelle Analogique - Fatigue, Échelle Visuelle Analogique - Anxiété) de l'état des sujets.

Prochaine partie

La suite de ce manuscrit tire pleinement parti de la tâche de lecture à voix haute de la base TILE pour concevoir quatre familles de descripteurs vocaux de la somnolence.

Bibliographie de la partie

Åkerstedt, T., et Gillberg, M. (**1990**). "Subjective and objective sleepiness in the active individual.," Int J Neurosci **52**, 29–37, doi: 10.3109/00207459008994241.

Aldrich, M. S., Chervin, R. D., et Malow, B. A. (**1997**). "Value of the multiple sleep latency test (MSLT) for the diagnosis of narcolepsy," Sleep **20**(8), 620–629, doi: 10.1093/sleep/20.8.620.

Aloshban, N., Esposito, A., et Vinciarelli, A. (**2021**). "What You Say or How You Say It? Depression Detection Through Joint Modeling of Linguistic and Acoustic Aspects of Speech," Cognitive Computation doi: 10.1007/s12559-020-09808-3.

Arand, D., Bonnet, M., Hurwitz, T., Mitler, M., Rosa, R., et Sangal, R. B. (**2005**). "The Clinical Use of the MSLT and MWT," SLEEP **28**(1), 123–144, doi: 10.1093/sleep/28.1.123.

Asaridou, S. S., et McQueen, J. M. (**2013**). "Speech and music shape the listening brain : evidence for shared domain-general mechanisms," Frontiers in Psychology **4**, doi: 10.3389/fpsyg.2013.00321.

Aydin Sayilan, A., Kulakaç, N., et Sayilan, S. (**2020**). "The effects of noise levels on pain, anxiety, and sleep in patients," Nursing in Critical Care nicc.12525, doi: 10.1111/nicc.12525.

Baird, A., Cummins, N., Schnieder, S., Krajewski, J., et Schuller, B. W. (**2020**). "An Evaluation of the Effect of Anxiety on Speech — Computational Prediction of Anxiety from Sustained Vowels," dans *Interspeech 2020*, pp. 4951–4955, doi: 10.21437/Interspeech.2020-1801.

Bastien, C. H., Vallières, A., et Morin, C. M. (**2001**). "Validation of the Insomnia Severity Index as an outcome measure for insomnia research," Sleep Medicine **2**(4), 297–307, doi: 10.1016/S1389-9457(00)00065-4.

Ben-Noun, L. L., Sohar, E., et Laor, A. (**2001**). "Neck Circumference as a Simple Screening Measure for Identifying Overweight and Obese Patients," Obesity Research **9**(8), 470–477, doi: 10.1038/oby.2001.61.

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., et Kupfer, D. J. (**1989**). "The Pittsburgh sleep quality index : A new instrument for psychiatric practice and research," Psychiatry Research **28**(2), 193–213, doi: 10.1016/0165-1781(89)90047-4.

Caraty, M.-J., et Montacié, C. (**2014**). "Vocal fatigue induced by prolonged oral reading : Analysis and detection," Computer Speech & Language 453–466.

Carskadon, M. A., et Dement, W. C. (**1977**). "Sleep tendency : an objective measure of sleep loss.," Sleep Research **6**(200), 940.

Chai, T., et Draxler, R. R. (**2014**). "Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature," Geoscientific Model Development **7**(3), 1247–1250, doi: 10.5194/gmd-7-1247-2014.

Cirillo, D., Catuara-Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S., et Mavridis, N. (**2020**). "Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare," npj Digital Medicine **3**(1), 81, doi: 10.1038/s41746-020-0288-5.

Cooke, M., et García Lecumberri, M. L. (**2021**). "How reliable are online speech intelligibility studies with known listener cohorts ?," The Journal of the Acoustical Society of America **150**(2), 1390–1401, doi: 10.1121/10.0005880.

Courvoisier, D., et Etter, J.-F. (**2008**). "Using item response theory to study the convergent and discriminant validity of three questionnaires measuring cigarette dependence.," Psychology of Addictive Behaviors **22**(3), 391–401.

Cummins, N., Baird, A., et Schuller, B. (**2018**). "Speech analysis for health : Current state-of-the-art and the increasing impact of deep learning," Health Informatics and Translational Data Analytics **151**, 1–54, doi: 10.1016/j.ymeth.2018.07.007.

Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., et Quatieri, T. F. (**2015**). "A review of depression and suicide risk assessment using speech analysis," Speech Communication **71**, 10–49, doi: https://doi.org/10.1016/j.specom.2015.03.004.

Di, Y., Wang, J., Li, W., et Zhu, T. (**2021**). "Using i-vectors from voice features to identify major depressive disorder," Journal of Affective Disorders **288**, 161–166, doi: 10.1016/j.jad.2021.04.004.

Doghramji, K., Mitler, M. M., Sangal, R. B., Shapiro, C., Taylor, S., Walsleben, J., Belisle, C., Erman, M. K., Hayduk, R., Hosn, R., O'Malley, E. B., Sangal, J. M., Schutte, S. L., et Youakim, J. M. (**1997**). "A normative study of the maintenance of wakefulness test (MWT)," Electroencephalography and Clinical Neurophysiology **103**(5), 554–562, doi: 10.1016/s0013-4694(97)00010-2.

Espy-Wilson, C., Lammert, A., Seneviratne, N., et Quatieri, T. (**2019**). "Assessing Neuromotor Coordination in Depression Using Inverted Vocal Tract Variables," dans *Interspeech 2019*, pp. 1448–1452.

Evangelista, E., Rassu, A. L., Barateau, L., Lopez, R., Chenini, S., Jaussent, I., et Dauvilliers, Y. (**2020**). "Characteristics associated with hypersomnia and excessive daytime sleepiness identified by extended polysomnography recording," Sleep doi: 10.1093/sleep/zsaa264.

Ewing, J. A. (**1984**). "Detecting Alcoholism : The CAGE Questionnaire," JAMA **252**(14), 1905.

Eyben, F., et Schuller, B. (**2015**). "Opensmile," ACM SIGMultimedia Records **6**, 4–13.

Golz, M., Sommer, D., Chen, M., Mandic, D., et Trutschel, U. (**2007**). "Feature Fusion for the Detection of Microsleep Events," Journal of VLSI Signal Processing **49**, 329–342.

Gosztolya, G. (**2019**). "Using Fisher Vector and Bag-of-Audio-Words Representations to Identify Styrian Dialects, Sleepiness, Baby & Orca Sounds," dans *Interspeech 2019*, pp. 2413–2417, doi: 10.21437/Interspeech.2019-1726.

Goy, H., Kathleen Pichora-Fuller, M., et van Lieshout, P. (**2016**). "Effects of age on speech and voice quality ratings," The Journal of the Acoustical Society of America **139**(4), 1648–1659, doi: 10.1121/1.4945094.

Guaita, M., Salamero, M., Vilaseca, I., Iranzo, A., Montserrat, J. M., Gaig, C., Embid, C., Romero, M., Serradell, M., León, C., de Pablo, J., et Santamaria, J. (**2015**). "The Barcelona Sleepiness Index : A New Instrument to Assess Excessive Daytime Sleepiness in Sleep Disordered Breathing," Journal of clinical sleep medicine **11**(11), 1289–1298, doi: 10.5664/jcsm.5188.

Hart, J. (**1981**). "Differential sensitivity to pitch distance, particularly in speech," The Journal of the Acoustical Society of America **69**(3), 811–821, doi: 10.1121/1.385592.

Helfer, K. S., et Freyman, R. L. (**2014**). "Stimulus and listener factors affecting age-related changes in competing speech perception," The Journal of the Acoustical Society of America **136**(2), 748–759, doi: 10.1121/1.4887463.

Hobson, D. E., Lang, A. E., Martin, W. R. W., Razmy, A., Rivest, J., et Fleming, J. (**2002**). "Excessive daytime sleepiness and sudden-onset sleep in Parkinson disease : a survey by the Canadian Movement Disorders Group," JAMA **287**(4), 455–463.

Huang, D.-Y., Tsao, Y., Chiori, H., et Kashioka, H. (**2011**). "Feature Normalization and Selection for Robust Speaker State Recognition," dans *IEEE - International Conference on Speech Database and Assessments*, doi: 10.1109/ICSDA.2011.6085988.

Huang, Z., Epps, J., Dale, J., et Chen, M. (**2018**). "Depression Detection from short Uterrances via Diverse Smartphones in Natural Environmental Conditions," dans *Interspeech 2018*.

Huckvale, M., Beke, A., et Ikushima, M. (**2020**). "Prediction of Sleepiness Ratings from Voice by Man and Machine," dans *Interspeech 2020*, doi: 10.21437/Interspeech.2020-1601.

Ihler, H. M., Meyrel, M., Hennion, V., Maruani, J., Gross, G., Geoffroy, P. A., Lagerberg, T. V., Melle, I., Bellivier, F., Scott, J., et Etain, B. (**2020**). "Misperception of sleep in bipolar disorder : an exploratory study using questionnaire versus actigraphy," International Journal of Bipolar Disorders **8**(1), 34, doi: 10.1186/s40345-020-00198-x.

Jaywant, A., et Pell, M. D. (**2010**). "Listener impressions of speakers with Parkinson's disease," Journal of the International Neuropsychological Society **16**(1), 49–57, doi: 10.1017/S1355617709990919.

Johns, M. W. (**1991**). "A New Method for Measuring Daytime Sleepiness : The Epworth Sleepiness Scale," Sleep **14**(6), 540–545, doi: https://doi.org/10.1093/sleep/14.6.540.

Joshi, A., Kale, S., Chandel, S., et Pal, D. K. (**2015**). "Likert scale : Explored and explained," British Journal of Applied Science & Technology **7**(4), 396, doi: 10.9734/BJAST/2015/14975.

Kaida, K., Takahashi, M., Åkerstedt, T., Nakata, A., Otsuka, Y., Haratani, T., et Fukasawa, K. (**2006**). "Validation of the Karolinska sleepiness scale against performance and EEG variables," Clinical Neurophysiology **117**(7), 1574–1581, doi: 10.1016/j.clinph.2006.03.011.

Kiss, G., et Vicsi, K. (**2017**). "Comparison of read and spontaneous speech in case of automatic detection of depression," dans *2017 8th IEEE International Conference on Cognitive Infocommunications (CogInfoCom)*, IEEE, Debrecen, pp. 000213–000218, doi: 10.1109/CogInfoCom.2017.8268245.

Krajewski, J., Batliner, A., et Golz, M. (**2009**). "Acoustic sleepiness detection : Framework and validation of a speech-adapted pattern recognition approach," Behavior Research Methods **41**(3), 795–804.

Kreiman, J., et Gerratt, B. R. (**2000**). "Sources of listener disagreement in voice quality assessment," The Journal of the Acoustical Society of America **108**(4), 1867–1876, doi: 10.1121/1.1289362.

Kreiman, J., et Gerratt, B. R. (**2005**). "Perception of aperiodicity in pathological voice," The Journal of the Acoustical Society of America **117**(4), 2201–2211, doi: 10.1121/1.1858351.

Kreiman, J., Gerratt, B. R., et Ito, M. (**2007**). "When and why listeners disagree in voice quality assessment tasks," The Journal of the Acoustical Society of America **122**(4), 2354–2364, doi: 10.1121/1.2770547.

Krupp, L. B., LaRocca, N. G., Muir-Nash, J., et Steinberg, A. D. (**1989**). "The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus," Archives of Neurology **46**(10), 1121–1123, doi: 10.1001/archneur.1989.00520460115022.

Littmann, M., Selig, K., Cohen-Lavi, L., Frank, Y., Hönigschmid, P., Kataka, E., Mösch, A., Qian, K., Ron, A., Schmid, S., Sorbie, A., Szlak, L., Dagan-Wiener, A., Ben-Tal, N., Niv, M. Y., Razansky, D., Schuller, B. W., Ankerst, D., Hertz, T., et Rost, B. (**2020**). "Validity of machine learning in biology and medicine increased through collaborations across fields of expertise," Nature Machine Intelligence **2**(1), 18–24, doi: 10.1038/s42256-019-0139-8.

Littner, M. R., Kushida, C., Wise, M., Davila, D. G., Morgenthaler, T., Lee-Chiong, T., Hirshkowitz, M., Loube, D. L., Bailey, D., Berry, R. B., Kapen, S., et Kramer, M. (**2005**). "Practice Parameters for Clinical Use of the Multiple Sleep Latency Test and the Maintenance of Wakefulness Test," Sleep **28**(1), 113–121, doi: 10.1093/sleep/28.1.113.

Ma, X., Yang, H., Chen, Q., Huang, D., et Wang, Y. (**2016**). "DepAudioNet : An Efficient Deep Model for Audio based Depression Classification," dans *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*, ACM Press, Amsterdam, The Netherlands, pp. 35–42, doi: 10.1145/2988257.2988267.

Maldonado, C. C., Bentley, A. J., et Mitchell, D. (**2004**). "A Pictorial Sleepiness Scale Based on Cartoon Faces," Sleep **27**(3), 541–548, doi: 10.1093/sleep/27.3.541.

Martin, V. P., Chapouthier, G., Rieant, M., Rouas, J.-L., et Philip, P. (**2020**a). "Using reading mistakes as features for sleepiness detection in speech," dans *Speech Prosody 2020*, Tokyo, Japan, pp. 985–989, doi: 10.21437/SpeechProsody.2020-201.

Martin, V. P., Rouas, J.-L., Micoulaud-Franchi, J.-A., et Philip, P. (**2020**b). "The Objective and Subjective Sleepiness Voice Corpora," dans *LREC 2020*, Marseille, France, pp. 6525–6533.

Martin, V. P., Rouas, J.-L., Micoulaud-Franchi, J.-A., Philip, P., et Krajewski, J. (**2021**). "How to Design a Relevant Corpus for Sleepiness Detection Through Voice ?," Frontiers in Digital Health **3**, 124, doi: 10.3389/fdgth.2021.686068.

Martin, V. P., Rouas, J.-L., et Philip, P. (**2020**c). "Détection de la somnolence dans la voix : nouveaux marqueurs et nouvelles stratégies," Traitement Automatique des Langues **61**(2), 67–90.

Martin, V. P., Rouas, J.-L., Thivel, P., et Krajewski, J. (**2019**). "Sleepiness detection on read speech using simple features," dans *10th Conference on Speech Technology and Human-Computer Dialogue*, Timisoara, Romania, doi: 10.1109/SPED.2019.8906577.

Matton, K., McInnis, M. G., et Mower Provost, E. (**2019**). "Into the Wild : Transitioning from Recognizing Mood in Clinical Interactions to Personal Conversations for Individuals with Bipolar Disorder," dans *Interspeech 2019*.

Miley, A. A., Kecklund, G., et Akerstedt, T. (**2016**). "Comparing two versions of the Karolinska Sleepiness Scale (KSS)," Sleep and Biological Rhythms **14**(3), 257–260, doi: 10.1007/s41105-016-0048-8.

Mitler, M. M., Gujavarty, K. S., et Browman, C. P. (**1982**). "Maintenance of wakefulness test : A polysomnographic technique for evaluating treatment efficacy in patients with excessive somnolence," Electroencephalography and Clinical Neurophysiology **53**(6), 658–661, doi: 10.1016/0013-4694(82)90142-0.

Nasir, M., Jati, A., Shivakumar, P. G., Nallan Chakravarthula, S., et Georgiou, P. (**2016**). "Multimodal and Multiresolution Depression Detection from Speech and Facial Landmark Features," dans *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge - AVEC '16*, ACM Press, Amsterdam, The Netherlands, pp. 43–50, doi: 10.1145/2988257.2988261.

Pan, W., Flint, J., Shenhav, L., Liu, T., Liu, M., Hu, B., et Zhu, T. (**2019**). "Re-examining the robustness of voice features in predicting depression : Compared with baseline of confounders," PLOS ONE **14**(6), e0218172, doi: 10.1371/journal.pone.0218172.

Preston, C. C., et Colman, A. M. (**2000**). "Optimal number of response categories in rating scales : reliability, validity, discriminating power, and respondent preferences," Acta Psychologica **104**(1), 1–15, doi: 10.1016/S0001-6918(99)00050-5.

Qian, K., Li, X., Li, H., Li, S., Li, W., Ning, Z., Yu, S., Hou, L., Tang, G., Lu, J., Li, F., Duan, S., Du, C., Cheng, Y., Wang, Y., Gan, L., Yamamoto, Y., et Schuller, B. W. (**2020**). "Computer Audition for Healthcare : Opportunities and Challenges," Frontiers in Digital Health **2**, 5, doi: 10.3389/fdgth.2020.00005.

Rhemtulla, M., Brosseau-Liard, P., et Savalei, V. (**2012**). "When can categorical variables be treated as continuous ? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions.," Psychological Methods **17**(3), 354–373, doi: 10.1037/a0029315.

Rutowski, T., Harati, A., Lu, Y., et Shriberg, E. (**2019**). "Optimizing Speech-Input Length for Speaker-Independent Depression Classification," dans *Interspeech 2019*, pp. 3023–3027, doi: 10.21437/Interspeech.2019-3095.

Sagaspe, P., Taillard, J., Chaumet, G., Guilleminault, C., Coste, O., Moore, N., Bioulac, B., et Philip, P. (**2007**). "Maintenance of Wakefulness Test as a Predictor of Driving Performance in Patients With Untreated Obstructive Sleep Apnea," Sleep **30**(3), 327–330, doi: 10.1093/sleep/30.3.327.

Sangal, R. B. (**1999**). "Subjective sleepiness ratings (Epworth sleepiness scale) do not reflect the same parameter of sleepiness as objective sleepiness (maintenance of wakefulness test) in patients with narcolepsy," Clinical Neurophysiology **110**(12), 2131–2135, doi: 10.1016/S1388-2457(99)00167-4.

Sato, M. (**2020**). "The neurobiology of sex differences during language processing in healthy adults : A systematic review and a meta-analysis," Neuropsychologia **140**, 107404, doi: 10.1016/j.neuropsychologia.2020.107404.

Schnack, H. G., et Kahn, R. S. (**2016**). "Detecting Neuroimaging Biomarkers for Psychiatric Disorders : Sample Size Matters," Frontiers in Psychiatry **7**, doi: 10.3389/fpsyt.2016.00050.

Schuller, B., Batliner, A., Bergler, C., Pokorny, F. B., Krajewski, J., Cychocz, M., Vollman, R., Roelen, S.-D., Schnieder, S., Bergelson, E., Cristia, A., Seidl, A., Warlaumont, A., Yankowitz, L., Nöth, E., Amiriparian, S., Hantke, S., et Schmitt, M. (**2019**). "The INTERSPEECH 2019 Computational Paralinguistics Challenge : Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," dans *Interspeech 2019*, doi: 10.21437/Interspeech.2019-1122.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., et Krajewski, J. (**2011**). "The INTERSPEECH 2011 Speaker State Challenge," dans *Interspeech 2011*, pp. 3201–3204, doi: 10.1.1.364.4935.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J., Weninger, F., et Eyben, F. (**2013**). "Medium-term speaker states-A review on intoxication, sleepiness and the first challenge," Comput. Speech Lang. **28**(2), 346–374.

Schweitzer, J. B., Cummins, T. K., et Kant, C. A. (**2001**). "Attention-deficit/hyperactivity disorder," The Medical Clinics of North America **85**(3), 757–777.

Shahid, A., Chung, S., Maresky, L., Danish, A., Bingeliene, A., Shen, J., et Shapiro, C. (**2016**). "The Toronto Hospital Alertness Test scale : relationship to daytime sleepiness, fatigue, and symptoms of depression and anxiety," Nature and Science of Sleep 41, doi: 10.2147/NSS.S91928.

Sparrow, A. R., LaJambe, C. M., et Van Dongen, H. P. (**2019**). "Drowsiness measures for commercial motor vehicle operations," Accident Analysis & Prevention **126**, 146–159.

Starke, G., De Clercq, E., Borgwardt, S., et Elger, B. S. (**2020**). "Computing schizophrenia : ethical challenges for machine learning in psychiatry," Psychological Medicine 1–7, doi: 10.1017/S0033291720001683.

Stasak, B., Epps, J., et Lawson, A. (**2018**). "Pathologic Speech and Automatic Analysis for Healthcare Applications (Batteries Not Included ?)," dans *17th Speech Science and Technology Conference (SST 2018)*.

Sturm, B. L. (**2016**). "Revisiting Priorities : Improving MIR Evaluation Practices," dans *Proceedings of the 17th International Society for Music Information Retrieval Conference*.

Sturm, B. L., Bardeli, R., Langlois, T., et Emiya, V. (**2014**). "Formalizing The Problem Of Music Description," dans *Proceedings of the 15th International Society for Music Information Retrieval Conference*.

Sussman, J. E., et Tjaden, K. (**2012**). "Perceptual Measures of Speech From Individuals With Parkinson's Disease and Multiple Sclerosis : Intelligibility and Beyond," Journal of Speech, Language, and Hearing Research **55**(4), 1208–1219, doi: 10.1044/1092-4388(2011/11-0048).

Thompson, W. F., Schellenberg, E. G., et Husain, G. (**2004**). "Decoding speech prosody : Do music lessons help ?," Emotion **4**(1), 46–64, doi: 10.1037/1528-3542.4.1.46.

Tremblay, P., Brisson, V., et Deschamps, I. (**2021**). "Brain aging and speech perception : Effects of background noise and talker variability," NeuroImage **227**, 117675, doi: 10.1016/j.neuroimage.2020.117675.

Vasquez-Correa, J., Arias-Vergara, T., Schuster, M., Orozco-Arroyave, J., et Nöth, E. (**2020**). "Parallel Representation Learning for the Classification of Pathological Speech : Studies on Parkinson's Disease and Cleft Lip and Palate," Speech Communication **122**, 56–67, doi: 10.1016/j.specom.2020.07.005.

Vasquez-Correa, J. C., Arias-Vergara, T., Klumpp, P., Perez-Toro, P. A., Orozco-Arroyave, J. R., et Nöth, E. (**2021**). "End-2-End Modeling of Speech and Gait from Patients with Parkinson's Disease : Comparison Between High Quality Vs. Smartphone Data," dans *ICASSP 2021*, pp. 7298–7302, doi: 10.1109/ICASSP39728.2021.9414729.

Verdot, C., Torres, M., Salanve, B., et Deschamps, V. (**2017**). "Children and adults body mass index in France in 2015. Results of the ESTEBAN study and trends since 2006," Bulletin Epidemiologique Hebdomadaire **13**, 234–241.

Weaver, T. E., Laizner, A. M., Evans, L. K., Maislin, G., Chugh, D. K., Lyon, K., Smith, P. L., Schwartz, A. R., Redline, S., Pack, A. I., et others (**1997**). "An instrument to measure functional status outcomes for disorders of excessive sleepiness," Sleep **20**(10), 835–843.

Willmott, C., et Matsuura, K. (**2005**). "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," Climate Research **30**, 79–82, doi: 10.3354/cr030079.

Yoho, S. E., Borrie, S. A., Barrett, T. S., et Whittaker, D. B. (**2019**). "Are there sex effects for speech intelligibility in American English ? Examining the influence of talker, listener, and methodology," Attention, Perception, & Psychophysics **81**(2), 558–570, doi: 10.3758/s13414-018-1635-3.

Zigmond, A. S., et Snaith, R. P. (**1983**). "The hospital anxiety and depression scale," Acta Psychiatrica Scandinavica **67**(6), 361–370, doi: 10.1111/j.1600-0447.1983.tb09716.x.

Quatrième partie

Descripteurs vocaux de la somnolence

Résumé

Cette partie introduit quatre familles de descripteurs extraits du signal de parole permettant de caractériser des sujets en train de lire un texte à voix haute.

Dans le chapitre 11, nous décrivons les marqueurs acoustiques de la voix préexistants à nos travaux et un nouvel ensemble de descripteurs de la qualité acoustique de la voix, validés avec des médecins. Ces marqueurs sont ensuite utilisés dans un système de classification automatique appliqué au *Sleep Language Corpus*.

Le chapitre 12 étudie l'utilisation de ces mêmes marqueurs pour la détection de la somnolence subjective instantanée, puis de la somnolence objective au long cours, sur la base TILE.

Dans le chapitre 13, nous proposons un nouveau type de marqueurs, mesurant une autre dimension de l'influence de la somnolence sur la lecture à voix haute : les erreurs de lecture.

Le chapitre 14 propose une automatisation des erreurs de lecture en étudiant les erreurs faites par des systèmes de transcription automatique.

Enfin, nous proposons dans le chapitre 15 un ensemble de descripteurs saisissant une troisième dimension de la parole lue : les durées et emplacements des pauses de lecture.

Mots-clés

Descripteurs acoustiques de la voix ; Erreurs de lecture ; Systèmes de transcription automatique ; Pauses de lecture

Publications associées

Martin, V. P., Rouas, J.-L., Thivel, P., et Krajewski, J. (2019). Sleepiness detection on read speech using simple features. *SPED 2019*.
https://doi.org/10.1109/SPED.2019.8906577

Martin, V. P., Chapouthier, G., Rieant, M., Rouas, J.-L., et Philip, P. (2020). Détection de la somnolence par estimation d'erreurs de lecture. *JEP 2020*, 1, 397-405.
https://aclanthology.org/2020.jeptalnrecital-jep.45.pdf

Martin, V. P., Chapouthier, G., Rieant, M., Rouas, J.-L., et Philip, P. (2020). Using reading mistakes as features for sleepiness detection in speech. *Speech Prosody 2020*, 985-989.
https://doi.org/10.21437/SpeechProsody.2020-201

Martin, V. P., Rouas, J.-L., Boyer, F., et Philip, P. (2021). Automatic Speech Recognition system errors for accident-prone sleepiness detection through voice. *EUSIPCO 2021*, 541-545.
https://doi.org/10.23919/EUSIPCO54536.2021.9616299

Martin, V. P., Rouas, J.-L., Boyer, F., et Philip, P. (2021). Automatic Speech Recognition systems errors for objective sleepiness detection through voice. *Interspeech 2021*, 2476-2480.
https://doi.org/10.21437/Interspeech.2021-291

Martin, V. P., Rouas, J.-L., et Philip, P. (2020). Détection de la somnolence dans la voix : Nouveaux marqueurs et nouvelles stratégies. *Traitement Automatique des Langues*, 61(2), 67-90.
https://atala.org/content/tal_61_2_3

Martin, V. P., Rouas, J.-L., et Philip, P. (2020). Détection de la somnolence objective dans la voix. *JEP 2020*, 1, 406-414.
https://aclanthology.org/2020.jeptalnrecital-jep.46.pdf

Martin, V. P., Arnaud, B., Rouas, J.-L., et Philip, P. (2022). Does sleepiness influence reading pauses in hypersomniac patients ? *Speech Prosody 2022*, 62-66.
https://doi.org/10.21437/SpeechProsody.2022-13

Martin, V. P., Rouas, J.-L., Basse, A., Caudron, B., Huillet, M., et Philip, P. (2022). Est-il possible d'annoter la naturalité des pauses lors de la lecture d'un texte à haute voix ? *JEP 2022*.

Chapitre 11

Marqueurs acoustiques de la somnolence subjective instantanée

Sommaire

Les travaux présentés dans ce chapitre ont été en partie menés dans le cadre du stage de Pierre Thivel, accueilli à l'occasion de son Projet de Fin d'Études pour l'École Nationale Supérieure de l'Électronique et de ses Applications (ENSEA) en 2019.

11.1 Contexte et motivations

La première étude portant sur l'influence de la somnolence ou de la fatigue sur les paramètres vocaux a été publiée en 1996 (Whitmore et Fisher, 1996). Dans cette étude, 12 soldats opérant habituellement dans des bombardiers ont été placés durant trois périodes de 36h dans des simulateurs de mission, chaque simulation étant séparée de 36h de repos. Lors de cette étude, les auteurs rapportent que la fréquence fondamentale et la durée des mots variaient au cours des missions, suivant les mêmes tendances que les tests cognitifs et les mesures subjectives de fatigue effectuées par les soldats.

Dès cette première étude, deux grandes familles de paramètres ont ainsi été évoquées : les paramètres acoustiques de la voix – fréquence, intensité, *qualité acoustique* – présentés dans ce chapitre, et des paramètres de qualité du discours – longueur des mots, fluidité, hésitations – qui feront les objets des chapitres 13, 14 et 15.

Ce chapitre présente les paramètres et systèmes de classification utilisés lors des deux compétitions internationales sur la détection et l'estimation de la somnolence dans la voix (section 11.2). Puis, dans la section 11.3 nous introduisons un nouvel ensemble de descripteurs acoustiques de la voix, validés avec des médecins du sommeil dans une approche interdisciplinaire. Enfin, nous proposons deux schémas de classification sur le *Sleepy Language Corpus*, basés respectivement sur un algorithme inspiré des moindres carrés (section 11.4) et sur un séparateur à vastes marges (section 11.5).

11.2 État de l'art

11.2.1 Compétition Interspeech 2011

La première compétition internationale sur la détection de la somnolence dans la voix a été proposée en 2011, au sein de la conférence internationale Interspeech 2011. Cette compétition, intitulée "The INTERSPEECH 2011 Speaker State Challenge" (Schuller *et coll.*, 2011) était divisée en deux parties : une pour la détection de l'état d'ébriété, et une pour la détection de la somnolence. Pour les deux tâches, à la fois les fichiers audios bruts et des marqueurs vocaux spécifiquement mis au point pour la compétition étaient proposés aux participants.

Description des marqueurs IS11

Les marqueurs les plus utilisés dans l'état de l'art pour la détection de la somnolence dans la voix sont ceux proposés lors de la compétition. Ils sont inclus dans la boîte à outils openSMILE (Eyben et Schuller, 2015) et sont composés de 60 descripteurs de bas niveau, sur lesquels sont calculées 39 fonctions. Les descripteurs de bas niveau et les fonctions sont présentés dans le tableau 11.1.

Marqueurs de bas niveau Les marqueurs de bas niveau inclus dans l'ensemble IS11 reposent sur trois types de marqueurs : des marqueurs dérivés de l'énergie, des marqueurs fréquentiels et des marqueurs calculés sur les parties vocaliques.

4 marqueurs d'énergie
▷ Somme du spectre auditif (*loudness*)
▷ Somme des coefficients RASTA
▷ Énergie RMS
▷ *Zero-Crossing Rate*
50 marqueurs fréquentiels
▷ 26 bandes du spectre RASTA
▷ MFCC 1–12
▷ Énergie spectrale 25–650 Hz et 1k–4kHz
▷ Point d'affaiblissement spectral à 25%, 50%, 75%, et 90%
▷ Flux spectral, Entropie, Variance, *Skewness*, *Kurtosis*, Pente
5 marqueurs vocaux
▷ F0
▷ Probabilité de voisement
▷ Jitter (local, delta)
▷ Shimmer (local)

33 fonctions de base
▷ Quartiles 1–3
▷ 3 intervalles interquartiles
▷ 1 % percentile (≈min)
▷ 99 % percentile (≈max)
▷ Intervalle de percentile 1%–99%
▷ Moyenne arithmétique, écart-type
▷ *skewness*, *kurtosis*
▷ Moyenne des distances entre les pics
▷ Écart-type de la distance entre les pics
▷ Moyenne des pics
▷ Moyenne des pics – moyenne arithmétique
▷ Pente de la régression linéaire et erreur quadratique
▷ Régression quadratique et erreur quadratique
▷ Centroïde du contour
▷ Durée du signal en dessous de 25%
▷ Durée du signal au-dessus de 90%
▷ Durée du signal pour lequel le marqueur diminue/augmente
▷ Gain de la prédiction linéaire
▷ Coefficients 1–5 de la prédiction linéaire
6 fonctions de F0
▷ Pourcentage d'échantillons non nuls
▷ Moyenne, max, min, écart-type de la longueur de l'échantillon
▷ Durée en seconde de l'entrée

Tableau 11.1 – Marqueurs de bas niveaux et fonctions statistiques qui leur sont appliquées pour définir les marqueurs vocaux IS11 (adapté de (Schuller *et coll.*, 2011)).

En plus de mesures directes telles que l'énergie RMS ou les statistiques fréquentielles (point d'affaiblissement, flux spectral, entropie ...), la majorité de ces descripteurs repose sur deux types de transformations du spectre fréquentiel très utilisées dans le traitement automatique du signal vocal : les MFCC – *Mel Frequency Cepstral Coefficients* (Oppenheim et Schafer, 2004) – et les coefficients RASTA-PLP (Hermansky et Morgan, 1994).

Composition des marqueurs acoustiques Les marqueurs acoustiques sont calculés à partir des marqueurs bas niveaux et des fonctions statistiques de la manière suivante :

— Les 33 fonctions de base du tableau de droite ainsi que la moyenne, le minimum, le maximum et l'écart-type de la longueur de segments sont appliquées sur chacun des 54 marqueurs de bas niveau mesurant l'énergie et la fréquence (tableau de gauche), et leur dérivée temporelle. Cela permet de définir 3996 marqueurs acoustiques de la voix ;

— Pour les cinq marqueurs vocaux de bas niveau et leurs dérivées temporelles, les fonctions de base ainsi que la moyenne quadratique et les durées d'augmentation et de

diminution du marqueur sont appliquées uniquement sur les régions vocaliques (probabilité de vocalisation supérieure à 0.7). Cela définit 360 marqueurs supplémentaires ;
— 12 marqueurs supplémentaires sont obtenus en appliquant les six fonctions de F0 sur le contour de F0 et sa dérivée temporelle, en incluant les parties non vocaliques pour lesquelles F0 est fixée à 0.

L'ensemble de descripteurs IS11 contient ainsi un total de 4368 marqueurs acoustiques.

Compétition

Le corpus SLC et les marqueurs IS11 ont été rendus publics en même temps, lors de la compétition précédemment mentionnée. Lors de cette compétition, la tâche proposée est un problème de classification binaire, prenant comme limite 7.5 sur la KSS mixte utilisée pour annoter le SLC (cf chapitre 6). Les classes étant déséquilibrées (34.5% de sujets somnolents), la métrique utilisée pour mesurer la performance des systèmes doit être adaptée (cf chapitre 16). Dans cette compétition, il s'agit du score de rappel non pondéré – UAR (*Unweighted Average Recall*) – défini comme la moyenne non pondérée du score de rappel sur chaque classe.

Lors de cette compétition, six systèmes ont été proposés. Un résumé des processus de sélection des marqueurs, des classifieurs employés et de leur performance sur la base de test du SLC est proposé dans le tableau 11.2.

Réf.	Sélection des marqueurs	Classification	UAR
(Schuller *et coll.*, 2011)†	-	SVM + SMOTE	70.3%
(Bozkurt *et coll.*, 2011)	RANSAC	SVM	65.4%
(Rodríguez, 2011)	LDA	UBM HMM semi continu à 32 états (phonèmes)	66.3%
(Montacié et Caraty, 2011)	WEKA : *SSFS* et *Best First*	SVM (M/F) sur 6 phonèmes, puis arbre de décision	69.4%
(Rahman *et coll.*, 2011)	test statistique du χ^2	SVM + SMOTE	71.0%
(Gajšek *et coll.*, 2011)	UBM-MAP	SVM	71.3%
(Huang *et coll.*, 2011)	ASIMPLS	ASIMPLS	**71.7%**

Tableau 11.2 – Systèmes proposés lors de la compétition sur la détection de la somnolence au court terme lors de la conférence internationale Interspeech 2011. ASIMPLS : *Asymetrical Statistically Inspired Modification of the Partial Least Squares algorithm*, HMM : *Hidden Markov Model*, *LDA : Linear Discriminant Analysis*, RANSAC : *Random Sampling Consensus*, SMOTE : *Synthetic Minority Over-Sampling TEchnique*, SSFS : *Subset Size Forward Selection*, SVM : *Support Vector Machine*, UBM : *Universal Background Model*.
† : référence de la compétition

Le vainqueur de la compétition (Huang *et coll.*, 2011) utilise un algorithme dérivé des moindres carrés – SIMPLS (DeJong, 1993) – pour la sélection des marqueurs acoustiques, et dépasse de 1.4% les performances de la référence avec un UAR de 71.7% : ce système est resté l'état de l'art pour la classification de la somnolence dans la voix avec le SLC jusqu'en 2019.

11.2.2 Compétition Interspeech 2019

Huit ans plus tard, en parallèle de la conférence Interspeech 2019, une autre compétition incluant l'estimation de la somnolence dans la voix a été proposée, intitulée "The INTERSPEECH 2019 Computational Paralinguistics Challenge : Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity". Parmi les quatre tâches proposées, l'une d'entre elles est

l'estimation du niveau de somnolence à partir du corpus SLEEP, que nous avons déjà introduit dans le chapitre 6.

Marqueurs acoustiques

Comme pour la compétition IS11, les marqueurs acoustiques présentés dans le précédent paragraphe étaient proposés aux participants. Ceux-ci étaient accompagnés de marqueurs issus de sacs de mots (*bag of words*), calculés sur les 65 marqueurs de bas niveau présentés précédemment et leur dérivée temporelle, le tout avec différentes tailles de dictionnaires. Enfin, ce challenge se déroulant en 2019, des marqueurs extraits par des techniques d'apprentissage profond sont proposés. Cet ensemble de marqueurs est obtenu grâce à des techniques d'apprentissage de représentation non supervisée avec des autoencodeurs séquence à séquence, implémentées grâce à la boîte à outils `AUDEEP` (Freitag *et coll.*, 2017).

Compétition

Le corpus SLEEP et les marqueurs précédemment cités ont été rendus publics en même temps, lors de la compétition IS19. Contrairement à la précédente compétition portant sur la détection de la somnolence – de façon binaire – le problème proposé dans celle-ci est l'estimation du niveau de somnolence du locuteur. Le but est ainsi de retrouver le score KSS associé à chaque échantillon, et la métrique de référence de la compétition est le coefficient de corrélation de Spearman.

Lors de cette compétition, six systèmes ont été proposés. Un résumé de ceux-ci et de leur performance sur la base de test du corpus SLEEP est proposé dans le tableau 11.3.

Résultats

Réf.	Système proposé	ρ
(Schuller *et coll.*, 2019)†	Moyenne des 3 SVR (un par catégorie de marqueurs)	0.343
(Elsner *et coll.*, 2019)	*Deep 1DCNN*	0.29
(Ravi *et coll.*, 2019)	*VQual features* + *elliptic enveloppe outlier detection* + SVR	0.331
(Wu *et coll.*, 2019b)	*Soft labelling* & *Ordinal Triplet Loss*	0.343
(Wu *et coll.*, 2019a)	vecteurs de Fischer + *Deep 2DCNN* + SVR	0.365
(Yeh *et coll.*, 2019)	BLSTM avec attention	0.369
(Gosztolya, 2019)	vecteurs de Fischer + *Bag of Words* + SVR	0.387

Tableau 11.3 – Systèmes proposés lors de la compétition sur l'estimation de la somnolence au court terme lors de la conférence internationale Interspeech 2019. *Bag of words* : sac de mots, *Deep 1DCNN* : réseaux profonds à base de neurones convolutifs à une dimension, *Deep 2DCNN* : réseaux profonds à base de neurones convolutifs à deux dimensions, SVR : *Support Vector Regressor*.
† : référence de la compétition

La nouveauté par rapport à la précédente compétition est l'utilisation de réseaux de neurones dans 3 des 5 systèmes proposés, avec différents niveaux de complexité. Ces approches conduisent cependant à des performances inférieures au vainqueur de la compétition, qui a utilisé des vecteurs de Fischer calculés sur le spectre des fichiers audio, puis un régresseur à vecteurs supports – SVR. Cette approche leur a permis d'établir l'état de l'art à 38.7% de corrélation entre le KSS estimé et la vérité terrain. Malgré de récents efforts basés sur des systèmes d'apprentissage profond (Fritsch *et coll.*, 2020; Amiriparian *et coll.*, 2020; Egas-Lopez

et Gosztolya, 2021), ceux-ci n'atteignent pas des performances à la hauteur de celles obtenues par le système proposé dans (Gosztolya, 2019).

11.3 Nouveaux marqueurs acoustiques (marqueurs personnalisés)

11.3.1 Limites des marqueurs précédents

Les marqueurs IS11 sont utilisés dans l'état de l'art et facilement reproductibles grâce à leur intégration dans la boîte à outils `openSMILE`. En revanche, ceux-ci souffrent d'un manque d'interprétabilité, qui est nécessaire lors d'une collaboration avec des médecins. En effet, ces marqueurs sont très variés : s'il est possible pour un ingénieur d'expliquer à un médecin ce qu'est la moyenne de l'énergie ou le *jitter* – voire ce qu'est un MFCC, des marqueurs comme le coefficient n°3 de la prédiction linéaire de la dérivée temporelle du 20e coefficient RASTA [1] sont difficilement interprétables et difficiles à relier à un comportement vocal particulier.

Nous avons donc utilisé pour nos études des marqueurs dont la définition est simple, dont le sens a été validé avec des médecins, dans une démarche de dialogue interdisciplinaire et d'explicabilité.

11.3.2 Description des marqueurs acoustiques

Tous les marqueurs acoustiques décrits dans les prochains paragraphes sont recensés dans le tableau 11.4. Ces marqueurs ont notamment été utilisés précédemment pour caractériser les modes phonatoires en voix chantée (Rouas et Ioannidis, 2016) ou la classification d'attitudes sociales (Rouas *et coll.*, 2019).

Statistiques concernant les parties voisées

Les marqueurs vocaux sont calculés en deux temps. Tout d'abord, nous extrayons d'une part les segments voisés grâce à l'algorithme d'extraction de la fréquence fondamentale ESPS de la boîte à outils `Snack` (Sjölander, 2004) ; et d'autre part les segments vocaliques grâce à la technique exposée dans (Pellegrino et Andre-Obrecht, 2000). Le premier sous-groupe de marqueurs est composé de statistiques sur ces segments, tandis que le second sous-groupe contient des marqueurs caractérisant la régularité de la production d'harmoniques sur les segments voisés. L'ensemble de ces marqueurs est ensuite moyenné pour obtenir un seul groupe de descripteurs acoustiques par échantillon.

Les statistiques obtenues sur les parties voisées et les parties vocaliques reflètent le comportement global du locuteur et sont les suivantes :

— la durée totale des parties voisées (en secondes) ;
— le pourcentage en durée des parties voisées ;
— la durée totale des segments vocaliques (en secondes) ;
— le pourcentage en durée des segments vocaliques.

Régularité de la production d'harmoniques sur les segments voisés

Une fois les parties voisées et les parties vocaliques extraites, nous mesurons la régularité de la production d'harmoniques sur ces segments grâce à des mesures de fréquence fondamentale et de courbes d'intensité. La tableau 11.4 résume les marqueurs ainsi conçus et leur

1. exemple réel inclus dans `openSMILE`

signification. Un exemple de variation de ces marqueurs au cours du temps avant le calcul de leur moyenne est proposé dans la figure 11.1.

— F0moy : la moyenne de la fréquence fondamentale sur les segments voisés ;
— F0var : la variance de la fréquence fondamentale sur les segments voisés ;
— F0pente : le coefficient directeur de l'approximation linéaire de la fréquence fondamentale sur un segment voisé ;
— F0max : le maximum de la fréquence fondamentale sur un segment voisé ;
— F0min : le minimum de la fréquence fondamentale sur un segment voisé ;
— F0E : l'étendue de la fréquence fondamentale sur un segment voisé (maximum - minimum).

Les mêmes paramètres sont calculés sur les courbes d'intensité (NRJmoy, NRJvar, NRJmax, NRJmin, NRJE). Il en résulte 12 paramètres vocaux supplémentaires (6 sur la fréquence fondamentale F0, 6 sur l'intensité). Nous avons également calculé les équivalents de F0moy, F0var, NRJmoy et NRJvar sur les segments vocaliques, ajoutant ainsi 4 paramètres vocaux à notre ensemble de descripteurs.

Cet ensemble de paramètres est complété par des paramètres reflétant la qualité acoustique des vocalisations, à travers des paramètres liés aux harmoniques et aux formants (cf chapitre 1) que nous avons calculés avec la boîte à outils Matlab Covarep (Degottex *et coll.*, 2014), qui a été modifiée pour les calculer seulement sur les segments voisés.

Nous complétons ainsi notre ensemble de paramètres avec l'amplitude des harmoniques (notées H1, H2, H4) et des statistiques sur les formants : amplitude (A1, A2, A3, A4), fréquence (F1, F2, F3, F4) et bande passante (B1, B2, B3, B4).

Nous calculons également la différence entre les amplitudes des harmoniques (H1-H2, H2-H4), la différence d'amplitude entre les harmoniques et les formants (H1-A1, H1-A2, H1-A3), la proéminence cepstrale (*Cepstral Peak Prominence* – CPP) et les rapports harmoniques sur bruit dans différentes plages de fréquences (HNR05 dans la plage [0,500Hz], HNR15 dans la plage [0,1500Hz], HNR25 dans la plage [0,2500Hz] et HNR35 dans la plage [0,3000Hz]). Tous ces paramètres sont moyennés sur chaque enregistrement, ce qui ajoute un total de 24 descripteurs.

NNous arrivons ainsi à un total de 44 paramètres vocaux.

11.4 Classification : ASIMPLS

11.4.1 Base de données – SLC

Cette section et la suivante présentent les résultats d'études menées sur le SLC (cf chapitre 6) dans le but de valider les marqueurs acoustiques présentés dans la section précédente, avant de pouvoir les utiliser sur la base TILE, qui était en cours de collecte au moment de cette étude. Pour cela, nous évaluons les performances d'algorithmes de classification identiques, basés sur chacun des deux ensemble de marqueurs : IS11, ou les Marqueurs Personnalisés, introduits dans la section précédente.

Dans cette première approche translationnelle, nous avons sélectionné uniquement les tâches de lecture du SLC. De plus, après la discussion menée au chapitre 9, nous sélections seulement les tâches de lecture dont la taille moyenne des échantillons est supérieure à 8 secondes : la lecture de la version en allemand de la fable *La bise et le soleil*) dont la durée moyenne est de 36.5 secondes ; la lecture de deux simulations de communication de trafic aérien (« flight1 » et « flight2 » de durées moyennes respectives de 9.7 secondes et 13.8 secondes) et la lecture d'une simulation de discours d'un contrôleur de trafic aérien « roger1 » (durée

	Nom des marqueurs
Fichier entier	**Statistiques des parties voisées** Nombre et pourcentage en durée des parties voisées **Statistiques des parties vocaliques** Nombre et pourcentage en durée des parties vocaliques
Segments voisés	**Statistiques sur F0** F0moy, F0var , F0 pente, F0max, F0min, F0E **Statistiques sur l'énergie** NRJmoy, NRJvar, NRJmax, NRJmin, NRJpente, NRJE **Statistiques sur les harmoniques et formants** amplitudes des harmoniques : H1, H2, H4 amplitudes des formants : A1, A2, A3, A4 fréquence des formants : F1, F2, F3, F4 bande passante des formants : B1, B2, B3, B4 différence d'amplitude des harmoniques : H1-H2, H2-H4 différence d'amplitude harmoniques / formants : H1-A1, H1-A2, H1-A3 **Cepstral Peak Prominence** **Rapport harmonique sur bruit (HNR)** HNR dans les plages 0-0.5kHz, 0-1.5kHz, 0-2.5kHz, 0-3.5kHz
Seg. vocaliques	**Statistiques sur F0** F0moy, F0var **Statistiques sur l'énergie** NRJmoy, NRJvar

TABLEAU 11.4 – Marqueurs acoustiques extraits utilisés dans les travaux présentés dans ce manuscrit.

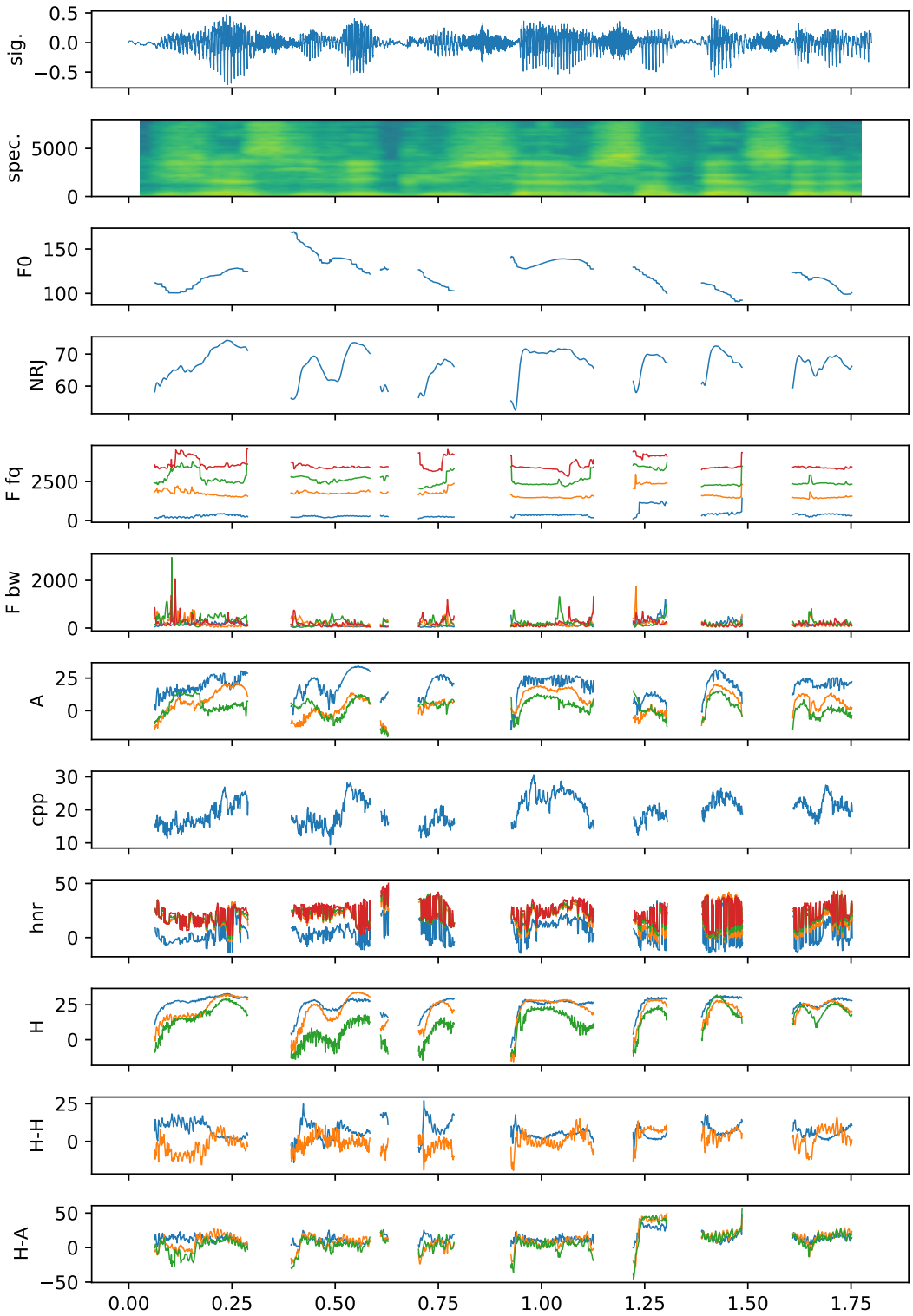

FIGURE 11.1 – Exemple d'évolution temporelle des marqueurs avant le calcul de leur moyenne sur l'ensemble de l'échantillon lors de la prononciation de la phrase « J'ai ainsi vécu seul, sans personne ... » Ceux-ci ne sont calculés que sur les parties voisées (c'est-à-dire là où f_0 est définie).

moyenne : 8.5 secondes). La fable est en allemand tandis que les trois autres tâches sont en anglais. Les statistiques de ce sous-corpus sont présentées dans le tableau 11.5.

Sexe	Classe	**Ent.**	**Dev.**	**Test**	**Total**
F	NS	10 loc. 109 éch. 4.15 (1.4) 24 min 23 s	8 loc. 88 éch. 4.0 (1.6) 18 m 43 s	9 loc. 73 éch. 4.18 (1.4) 17 m 48 s	27 loc. 270 éch. 4.11 (1.5) 1 h 54 s
	S	5 loc. 106 éch. 8.12 (0.5) 18 m 23 s	6 loc. 76 éch. 8.13 (0.7) 15 m 18 s	6 loc. 86 éch. 8.21 (0.9) 17 m 59 s	17 loc. 268 éch. 8.15 (0.7) 51 m 40 s
H	NS	10 loc. 54 éch. 4.8 (1.2) 17min 4s	7 loc. 27 éch. 3.9 (1.6) 8 m 17s	9 loc. 52 éch. 3.5 (1.8) 15 m 31 s	26 loc. 133 éch. 4.1 (1.6) 40 m 53s
	S	4 loc. 33 éch. 8.7 (0.9) 7 m 8 s	7 loc. 56 éch. 8.7 (1.0) 14 m 4s	3 loc. 30 éch. 8.14(0.9) 6 m 41 s	14 loc. 119 éch. 8.6 (1.0) 27 m 53 s
Total	NS	20 loc. 164 éch. 4.3 (1.4) 41 m 38 s	15 loc. 115 éch. 4.0 (1.6) 26 m 59 s	18 loc. 125 éch. 3.9 (1.6) 33 m 19 s	53 loc. 404 éch. 4.1 (1.5) 1 h 41 m 57s
	S	9 loc. 139 éch. 8.3 (0.7) 25 m 31 s	13 loc. 132 éch. 8.4 (0.9) 29 m 21 s	9 loc. 116 éch. 8.2 (0.9) 24 m 40 s	31 loc. 387 éch. 8.3 (0.8) 1 h 19 m 33s

TABLEAU 11.5 – Nombre de locuteurs, nombre d'échantillons, KSS moyenne (écart-type) et durée cumulée d'enregistrements du sous-corpus de la base SLC contenant uniquement des tâches de lecture. S : somnolent (KSS $\geq$ 7.0) ; NS : non somnolent (KSS $<$ 7.0) ; Ent. : entraînement ; Dev. : développement.

11.4.2 Méthode

Le système état-de-l'art sur la classification de la somnolence à court terme sur le SLC (Huang *et coll.*, 2011, 2014) repose sur une version asymétrique de l'adaptation statistique de l'algorithme des moindres carrés – ASIMPLS (*Asymetrical Statistically Inspired Modification of the Partial Least Squares*).

Principe

L'idée d'utiliser l'algorithme ASIMPLS vient du déséquilibre entre les classes S et NS du SLC (cf chapitre 6). Les techniques habituellement utilisées pour faire face à ce déséquilibre sont usuellement un suréchantillonnage de la classe minoritaire (c'est par exemple ce que fait l'algorithme SMOTE (Chawla *et coll.*, 2002)), en créant des échantillons selon une règle définie ; ou alors un sous-échantillonnage de la classe majoritaire, en ne travaillant qu'avec une partie des échantillons de celle-ci. Dans les deux cas, le but est de changer la distribution des échantillons, avant d'appliquer les méthodes de classification usuelles, qui fonctionnent mieux sur des distributions de classes équilibrées. Dans leur article pour la compétition IS11, Huang

et coll. (2011) cherchent à améliorer la classification des échantillons minoritaires sans changer le corpus fourni (SLC). Ils proposent pour cela l'algorithme suivant.

Algorithme des Moindres Carrés

L'ASIMPLS se base sur l'algorithme des moindres carrés, dont l'hypothèse est qu'il existe un vecteur de variables latentes **r** indépendant du locuteur, et deux matrices de transformation **Q** et **P**, dépendantes du locuteur, telles que :

$$\mathbf{X} = \mathbf{Q}\mathbf{r} + \mathbf{e}_x \tag{11.1}$$

$$\mathbf{y} = \mathbf{P}\mathbf{r} + \mathbf{e}_y \tag{11.2}$$

où $\mathbf{X} \in \mathbb{R}^{N \times M}$ et $\mathbf{y} \in \mathbb{R}^N$ sont respectivement la matrice des M marqueurs observés N fois et les labels correspondants, et $\mathbf{e}_x$ et $\mathbf{e}_y$ les résidus du modèle.

Il est possible de montrer que ce système se rapporte à une unique équation :

$$\mathbf{y} = \mathbf{B}\mathbf{X} + \mathbf{e} \tag{11.3}$$

où **B** est la matrice de régression dépendant de **P** et de **Q**, et **e** est le résidu de la régression.

Le seul hyperparamètre de cet algorithme est l, le rang de la matrice **B**, aussi appelé *nombre de composantes* : c'est cet hyperparamètre qui joue le rôle de seuil entre sous-apprentissage et surapprentissage. Si l est trop petit, le nombre de composantes choisies sera trop faible pour représenter la complexité du problème. Au contraire, si l est trop élevé, des composantes seront superflues à la généralisation du problème et risquent de participer à un surapprentissage des données.

Il existe de nombreux algorithmes pour la résolution de ce problème de régression des moindres carrés. Celle proposée par Huang *et coll.* (2014) repose sur l'algorithme SIMPLS, qui a l'avantage d'être performant en tout en évitant de coûteux calculs de matrices inverses.

Par la suite, trois phases sont différenciées :

1. Une phase d'entraînement, pour calculer les paramètres de l'algorithme ;
2. La détermination de $\mathbf{b}'$, un terme correctif, permettant de considérer des frontières de classe circulaires plutôt que linéaires ;
3. Une phase de test, permettant d'appliquer l'algorithme à la base de test et estimer les labels $\hat{\mathbf{y}}$ de ceux-ci.

Les étapes 1 et 2 du précédent algorithme permettent de calculer **m**, la matrice de projection vers l'espace des vecteurs de score, $\hat{\mathbf{T}}$, les scores estimés et $\mathbf{b}'$, le facteur correctif. L'équation d'inférence de la phase de test devient alors :

$$\hat{\mathbf{y}} = \text{signe}\left(\sum_{i=0}^{l-1} m_i \hat{\mathbf{t}}_i - \mathbf{b}'\right) \tag{11.4}$$

Les algorithmes complets des trois procédures ainsi que les détails concernant le terme correctif **b'** sont présentés en Annexe F.

11.4.3 Résultats

Nombre de composantes

Nous avons déterminé le seul hyperparamètre du système, l, en étudiant les performances du système en fonction des valeurs de celui-ci. Le système est entraîné sur la base d'entraînement du sous-corpus du SLC et les performances sont évaluées sur la base de développement. Comme pour le challenge IS11, la mesure de performances utilisée est l'UAR.

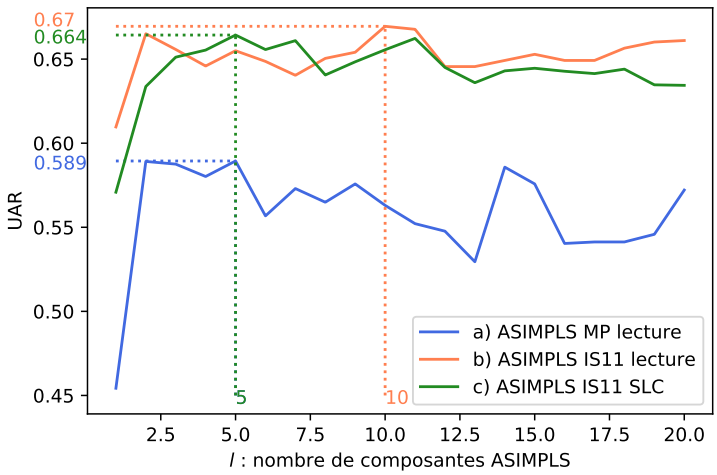

FIGURE 11.2 – Performances de l'algorithme ASIMPLS sur la base de développement en fonction du nombre de composantes sélectionnées. *MP* : marqueurs personnalisés. *IS11* : marqueurs acoustiques proposés pour IS11. *lecture* : sous-corpus du SLC ne contenant que les tâches de lecture.

Pour le système (a), appliquant l'algorithme aux marqueurs personnalisés calculés sur le sous-corpus de lecture du SLC, le nombre optimal de composantes est 5, ce qui conduit à un UAR de 58.9%. Le l optimal est le même pour les marqueurs IS11 calculés sur le SLC entier (c), donnant 66.4% de performances. Au contraire, pour les marqueurs IS11 calculés sur le sous-corpus de lecture (b), le nombre optimal de marqueurs est de 10, permettant d'atteindre 67.0% d'UAR.

Performances

Les performances de ces trois systèmes sur la base de test sont reportées dans le tableau 11.6. Avec des UAR tous inférieurs à 70%, les trois systèmes testés ont des performances bien en deçà du système proposé par Huang *et coll.* (2014).

	Système	Corpus	l	Sen.	Spé.	UAR
(a)	MP	SLC lecture	5	44.3%	83.0%	63.7%
(b)	IS11	SLC lecture	10	56.8%	78.4%	67.6%
(c)	IS11	SLC	5	67.8%	68.8%	68.3%
(d)	IS9+IS10+IS11 (Huang *et coll.*, 2011)	SLC	n.c.	64.3%	79.1%	**71.7%**

TABLEAU 11.6 – Performances de l'algorithme d'ASIMPLS sur le SLC et sur le sous-corpus de tâches de lecture, en fonction de marqueurs utilisés (MP : marqueurs personnalisés, IS11 : Interspeech 2011). Classification binaire entre la classe S (KSS $\geq$ 7.5) et la classe NS (KSS $<$ 7.5). n.c. : non communiqué dans l'article ; Sen. : Sensibilité ; Spé. : spécificité.

À la différence des systèmes que nous avons implémentés, le système proposé par Huang *et coll.* (2014) fait la fusion entre trois systèmes d'ASIMPLS, chacun basé sur un ensemble de marqueurs acoustiques différents (IS9, IS10 et IS11, tous disponibles dans openSMILE). Par ailleurs, ni l'article de 2011 ni celui de 2014 ne précisent les hyperparamètres sélectionnés pour la classification, rendant la reproduction des résultats difficile.

En conclusion, l'algorithme ASIMPLS ne semble pas convenir à notre volonté de réduire le nombre de marqueurs acoustiques et de se concentrer sur ceux qui sont pertinents pour des tâches de lecture. Nous avons donc mis au point un autre schéma de classification, basé sur le classifieur le plus employé dans l'état de l'art : les séparateurs à vastes marges (*Support Vector Machine* – SVM).

11.5 Classification : SVM

11.5.1 Méthode

Principe

La majorité des systèmes proposés pour les compétitions IS11 et IS19 – systèmes de référence compris – utilisent un SVM ou un SVR pour l'étape de classification ou de régression, et avec différents algorithmes pour la phase de sélection des marqueurs ou avec des marqueurs de leur conception. Le classifieur SVM a été choisi à la fois en raison de sa capacité à classifier de petites bases de données grâce à des ensembles de descripteurs de grandes dimensions, mais aussi pour « l'astuce du noyau »[2], qui permet de calculer des frontières de classes avec d'autres géométries qu'uniquement linéaires. Dans ce schéma de classification, nous proposons de sélectionner uniquement les N marqueurs les plus corrélés avec la vérité terrain avant la classification par le SVM. Par ailleurs, nous étudions l'influence de centrer les marqueurs de chaque locuteur.

Schéma de classification

En raison du faible nombre d'échantillons et de leur déséquilibre entre les deux classes S et NS, un suréchantillonnage de la classe minoritaire est appliqué grâce à l'algorithme SMOTE – *Synthetic Minority Over-sampling Technique* (Chawla *et coll.*, 2002) implémenté dans la boîte à outils Python `Sklearn` (Pedregosa *et coll.*, 2011).

Le schéma de classification proposé est représenté dans la figure 11.3 et se décompose de la manière suivante :

1. Centrage des paramètres vocaux par locuteur. En soustrayant la moyenne des marqueurs vocaux d'un locuteur à tous les marqueurs de ce sujet, nous éliminons les facteurs propres au locuteur (sexe, âge, physionomie des voies respiratoires…) et nous gardons uniquement les variations instantanées des paramètres vocaux, qui ne sont plus pollués par les marqueurs traits s'exprimant dans la voix. Cette méthodologie semble d'autant plus pertinente que l'on cherche à estimer la somnolence subjective à court terme et non un état général de somnolence sur le long terme du locuteur ;
2. Calcul pour chaque marqueur vocal de la corrélation (ρ de Spearman) entre le marqueur et la mesure de somnolence (KSS). Cela permet d'ordonner les marqueurs vocaux du plus corrélé au moins corrélé avec la mesure de somnolence. Ce calcul se fait sur les ensembles d'entraînement et de développement ;
3. Sélection du nombre de marqueurs et des paramètres optimaux du classifieur. Pour cela, nous calculons les performances du système (sur la base d'entraînement vs la base de développement) pour les 1, 2, …, 44 marqueurs vocaux précédemment triés, et nous conservons le nombre de marqueurs vocaux et les paramètres du classifieur

2. *kernel trick*

fournissant les meilleures performances. Le classifieur utilisé est un SVM, dont les paramètres sont le type de noyau (linéaire ou gaussien), et les paramètres C et γ.

4. Les paramètres C et γ obtenus lors de l'étape 3 sont utilisés pour entraîner le SVM sur les sous-corpus entraînement et développement. Nous inférons ensuite les classes de somnolence estimées de chaque échantillon du sous-corpus de test.

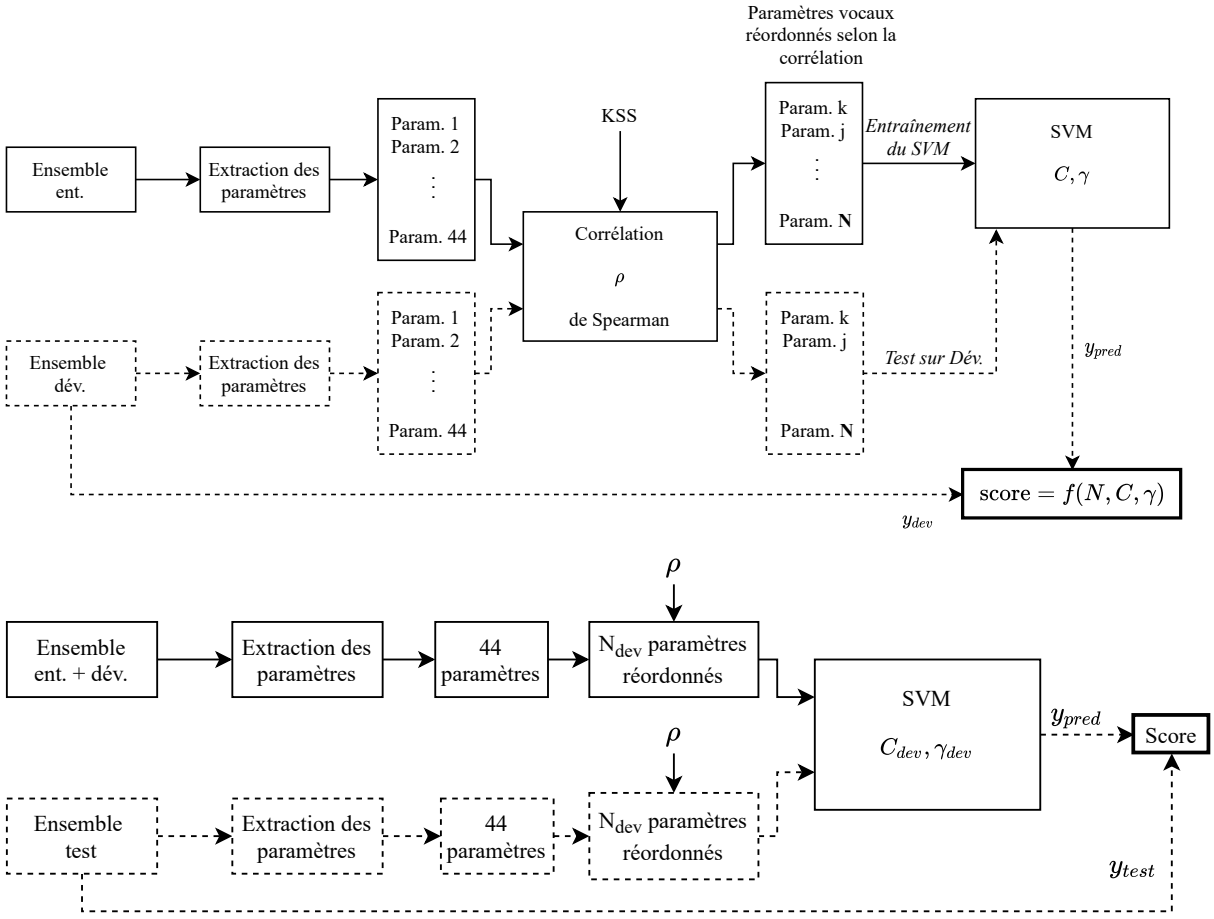

FIGURE 11.3 – Schéma explicatif de la classification proposée en prenant l'exemple de 44 marqueurs personalisés.

11.5.2 Marqueurs sélectionnés

Nombre de marqueurs

De même que pour l'ASIMPLS, nous traçons dans la figure 11.4 les performances sur la base de développement en fonction du nombre de marqueurs sélectionnés, pour les marqueurs IS11 (système e) et les marqueurs personnalisés (système f). Pour les marqueurs personnalisés, les 23 marqueurs corrélant le plus avec la KSS permettent d'obtenir un UAR sur la base de développement de 68.1%, tandis que pour les marqueurs IS11, les 101 marqueurs corrélant le plus avec la KSS permettent d'obtenir un UAR de 70.3%. Ce sont donc ces marqueurs qui seront employés pour la classification.

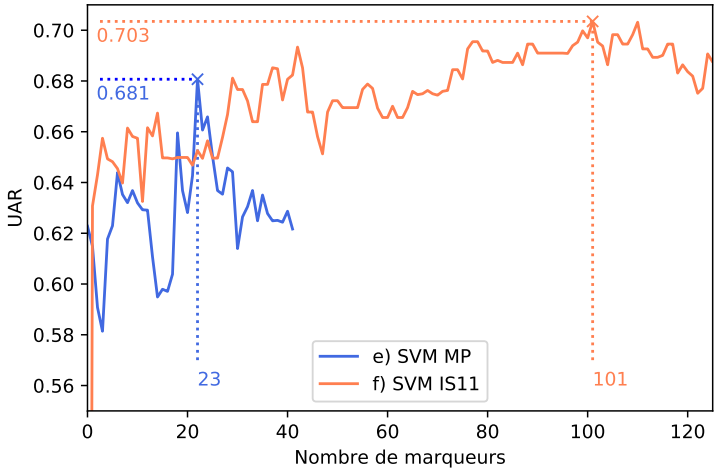

FIGURE 11.4 – Performances sur la base de développement des systèmes basés sur un SVM en fonction du nombre de marqueurs acoustiques (IS11 ou MP) sélectionnés. MP : Marqueurs Personalisés.

11.5.3 Performances

Les performances de tous les systèmes utilisant la méthode basée sur un SVM sont rassemblées dans le tableau 11.7.

	Marqueurs	#Marqueurs	Centrage	Corpus	limKSS	UAR
(e)	MP	23	oui	SLC lecture	7.5	**76.4%**
(f)	IS11	101	oui	SLC lecture	7.5	68.8%
(g)	MP	23	oui	SLC	7.5	66.8%
(h)	MP	23	oui	SLC lecture	7.0	**77.6%**
(i)	MP	23	non	SLC lecture	7.0	66.8%
(j)	MP	23	oui	SLC lecture filtré	7.0	72.4%

TABLEAU 11.7 – Paramètres et performances des systèmes basés sur un SVM.

Comparaison avec l'ASIMPLS

Les systèmes (e), (f) et (g) proviennent de l'étude sur les tâches de lecture du SLC menée dans (Martin *et coll.*, 2019), pour laquelle la limite de la KSS séparant les deux classes S et NS est de 7.5. Le système (e), utilisant 23 marqueurs acoustiques atteint un UAR de 76.4%, ce qui représente un gain de 4.7% (en UAR absolu) par rapport à l'état de l'art, qui atteint 71.7%. En revanche, le système (f) basé sur les marqueurs IS11 atteint un UAR de 68.8%, qui est inférieur à l'état de l'art de presque 3% : les marqueurs sélectionnés semblent particulièrement pertinents pour cette tâche de classification. De plus, la comparaison des systèmes (e) et (g) permet de mettre en lumière que les marqueurs sélectionnés par le système (e) sont spécifiques des tâches de lectures, et ne sont pas adaptés aux autres tâches du SLC. Les performances du système (e) sont significativement supérieures à celles du système ASIMPLS équivalent

(+12.7% d'UAR absolus entre les systèmes a et e).

Autre seuil de KSS pour les classes S et NS

Afin d'assurer une comparaison valide avec la base TILE, nous avons considéré le même schéma de classification en changeant la limite de la KSS délimitant les classes S et NS. En effet, alors que dans le SLC la vérité terrain est la moyenne de l'annotation du sujet et deux annotateurs externes, conduisant à un nombre réel, la KSS de la base TILE n'est annotée que par le patient lui-même. Cela se traduit par une vérité terrain qui ne contient que des valeurs entières. Nous choisissons donc la valeur « 7 - Somnolent, mais sans effort pour rester éveiller », à la fois la plus proche de l'ancienne limite de 7.5 et qui mesure bien la somnolence, et non l'effort d'éveil comme stipulé par le niveau « 8 - Somnolence, mais avec des efforts pour rester éveiller ». Ainsi, en conservant les 23 précédents marqueurs vocaux et en réentraînant le classifieur avec cette nouvelle limite, le système (h) permet d'obtenir 77.6% d'UAR sur les tâches de lecture du SLC, dépassant la précédente meilleure performance de 76.4% atteinte par le système (e).

De plus, la comparaison entre le système (h) et le système (i) montre l'importance du centrage pour ces marqueurs : sans ce dernier, la pollution des marqueurs vocaux par les traits des locuteurs rend plus difficile la tâche de classification. Le classifieur proposé atteint alors un UAR de 66.8%, soit plus de 10% de moins que la version avec centrage des marqueurs.

Locuteurs surreprésentés

Les avertissements concernant une surreprésentation de certains locuteurs dans le SLC soulevés dans le chapitre 8 nous incitent à vérifier si les résultats que nous avons présentés ne sont pas influencés par ce biais.

Dans le sous-corpus du SLC considéré, chaque locuteur est enregistré en moyenne sur 13 enregistrements (nombre moyen d'enregistrements par sujet : 13.4, é-t : 19.4). Sur les 94 sujets enregistrés, 5 ont produit à eux seuls 359 échantillons des 791 compris dans le sous-corpus de lecture du SLC. Deux d'entre eux sont dans la base d'entraînement (n°38 et n°39), enregistrés respectivement sur 56 et 95 échantillons. Deux autres comptant 36 et 75 échantillons sont dans la base de développement (n°40 et n°41) tandis que le dernier (n°42) est dans la base de test et compte un total de 96 échantillons. Ces cinq locuteurs sont ceux qui avaient déjà été identifiés comme intrus dans le chapitre 8.

Nous faisons l'hypothèse que le très grand nombre d'échantillons par locuteur dans ce sous-corpus permet une meilleure estimation des traits vocaux propres au locuteur, et donc un meilleur centrage lors de l'étape 1 du schéma de classification. Pour vérifier cette hypothèse, nous avons réappliqué le même schéma de classification que le système (h), en excluant ces locuteurs. Ce système (j) conduit à un UAR de 72.4%, qui est inférieur de plus de 5% d'UAR absolus au système (h), mais reste supérieur à l'état de l'art de la compétition IS11 (d).

11.5.4 Discussions

Performances en fonction de la limite entre classes de somnolence

Dans le SLC, la limite de la KSS utilisée pour différencier les deux classes Somnolent et Non-Somnolent a été fixée à 7.5. Cependant, celle-ci correspond au cas très particulier de différenciation de performances des sujets, et non à une différence de niveau de somnolence subjective.

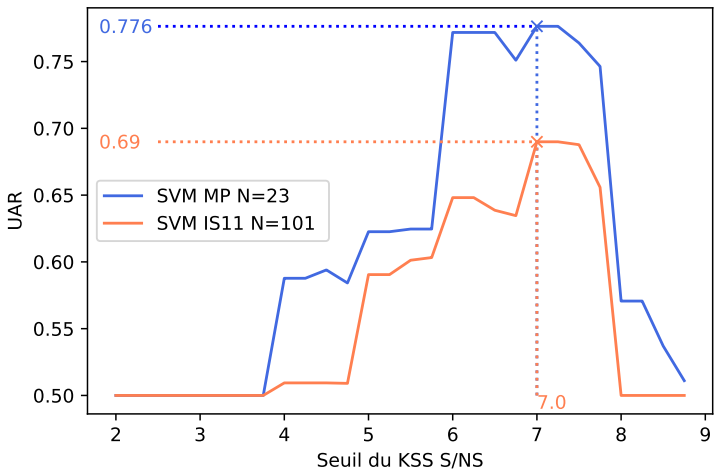

FIGURE 11.5 – Performances des systèmes (e) et (f) en fonction de la limite de la KSS utilisée pour discriminer les deux classes S et NS. MP : Marqueurs Personalisés, N : nombre de marqueurs acoustiques.

Cette échelle étant pluridimensionnelle, nous avons étudié dans la figure 11.5 les variations des performances du système (e) en faisant varier la limite délimitant les deux classes de la classification binaire. Pour chaque limite de KSS, les labels S et NS sont recalculés, le système (e) est réentraîné, et les labels de test sont réinférés.

Les meilleures performances sont ainsi obtenues pour une limite de 7.0, qui se trouve être le système (h). Le système atteint également de très bonnes performances pour des limites se situant entre 6 (inclus) et 8 (exclus), qui sont supérieures à 75% d'UAR. En dehors de cette zone, les performances de classification chutent brusquement sous les 65% d'UAR, rendant l'exploitation de tels systèmes impossible.

Les marqueurs identifiés par le système (e) sont donc spécifiques d'un comportement associé à une KSS mixte entre « 6 - Un peu somnolent » et « 8 - Somnolent, mais avec des efforts pour rester éveiller ». Nous faisons l'hypothèse que la même tendance aurait pu être observée pour l'item « 9 - Très somnolent, avec de grands efforts pour rester éveillé, luttant contre le sommeil » si les échantillons annotés avec un score supérieur à 9 avaient été plus nombreux dans le sous-corpus du SLC étudié. Ainsi, les marqueurs vocaux identifiés ne sont pas spécifiques des manifestations comportementales de la somnolence, mais bien de l'effort de maintien d'éveil.

Interprétation des marqueurs sélectionnés

L'une des plus grandes contraintes de ce travail était de sélectionner des caractéristiques permettant de relier les modifications physiologiques de la voix du patient à la somnolence. Ainsi, nous rapportons dans le tableau 11.8 une analyse des marqueurs sélectionnés par le système (e) au regard de la littérature précédente.

Tout d'abord, de manière similaire à (Dhupati *et coll.*, 2010), une augmentation des parties voisées et des voyelles est observée. Cette observation peut être un indice de l'augmentation des hésitations des locuteurs somnolents. La diminution des valeurs de F0moy, F0min, F0max, F0E, F0pente, de la fréquence F1 [également observées dans (Krajewski *et coll.*, 2009) et (Greeley *et coll.*, 2006)], de la largeur de bande passante de F1 et de l'amplitude des deuxième et

Marqueurs	ρ de Spearman	p	rang
durée des parties voisées	0.06	0.17	23
durée des voyelles	0.05	0.29	19
F0moy (voyelles)	-0.32	***	1
F0moy	-0.27	***	2
F0pente	-0.09	*	16
F0min	-0.20	***	4
F0max	-0.24	***	3
F0E	-0.08	0.06	17
NRJvar (voyelles)	-0.07	0.1	21
NRJvar	-0.07	0.1	22
NRJpente	0.13	**	10
NRJmin	0.14	***	8
NRJE	-0.16	***	6
H1	0.10	*	14
H2	0.13	**	9
A2	-0.08	0.07	18
A3	-0.10	*	13
F1	-0.19	***	5
B1	-0.10	*	15
H1A1	0.07	*	20
H1A2	0.12	**	11
H1A3	0.14	***	7
HNR05	-0.12	**	12

TABLEAU 11.8 – Description, corrélation avec la KSS et rang des 23 marqueurs sélectionnés par le système (e). $p < 0.05$: *, $p < 0.01$: **, $p < 0.001$: ***

troisième formants témoigne d'un déplacement des fréquences contenues dans la voix vers des valeurs plus basses. Cela concorde avec les résultats obtenus par Nwe *et coll.* (2006), Krajewski *et coll.* (2009) et McGlinchey *et coll.* (2011). De plus, la diminution des valeurs de F0E et F0pente sont des indices d'une réduction de l'étendue fréquentielle utilisée lors du processus vocal.

Contrairement aux observations faites dans (Krajewski *et coll.*, 2009), l'étendue de l'énergie, la valeur absolue de la pente de l'énergie et la variance de l'énergie diminuent avec la somnolence. Combinées à l'augmentation des basses fréquences, et une diminution des variations de l'énergie, ces observations expriment une diminution des nuances dans la prosodie des sujets somnolents. Nous supposons que la légère augmentation de la fréquence de la première harmonique, qui semble contraire aux observations précédentes, est due à la modification du flux d'air expiré qui modifie la distribution des harmoniques, mais pas des formants (Hillenbrand *et coll.*, 1994). Ceci est cohérent avec la diminution de l'HNR (HNR05 dans notre cas) également observée dans (Boyer *et coll.*, 2016).

Toutes ces observations conduisent à l'hypothèse que les locuteurs somnolents peinent à produire la même variété de nuances de fréquences, d'énergie et de qualité des parties voisées que leurs homologues éveillés.

11.6 Conclusion et perspectives

En conclusion, nous avons proposé un nouvel ensemble de marqueurs acoustiques que nous avons validés sur un sous-corpus du SLC, composé des enregistrements de tâches de lecture, afin de permettre une translation des résultats obtenus vers la base TILE. Ces nouveaux marqueurs acoustiques conçus dans une démarche d'explicabilité aux médecins ont été reliés au comportement de somnolence (tel que mesuré par la KSS dans le SLC).

L'approche utilisée par l'état de l'art – SIMPLS – est abandonnée au profit des SVM, qui obtient de meilleures performances et permet d'aligner le système proposé avec la majorité des systèmes de l'état de l'art.

Par ailleurs, nous avons proposé une étude de la robustesse des marqueurs sélectionnés et du classifieur à un changement de définition de la somnolence en faisant varier la valeur limite de la KSS différenciant les deux classes « Somnolent » et « Non Somnolent. » Cela a permis de montrer que le système (e) atteint de bonnes performances dans une large plage de somnolence (le système peut détecter à la fois les somnolences sévères et légères, suivant la limite avec laquelle il a été entraîné), permettant une flexibilité suivant les besoins des cliniciens ou des industries.

Enfin, nous avons pu observer une cohérence entre les marqueurs sélectionnés dans cette étude et la littérature précédemment publiée sur le sujet, confirmant la pertinence des descripteurs que nous avons proposés.

Après avoir étudié la somnolence subjective instantanée sur le SLC, le prochain chapitre étudie la pertinence de ces marqueurs acoustiques pour la détection de la somnolence subjective, mais aussi objective, telle que mesurée par les latences d'endormissement au TILE, en utilisant le corpus du même nom.

Chapitre 12

De la somnolence subjective instantanée à la somnolence objective au long cours

Sommaire

12.1 Motivations

Dans le chapitre précédent, nous avons introduit de nouveaux marqueurs acoustiques, que nous avons validés sur la tâche de détection des manifestations comportementales de la somnolence proposée dans le *Sleepy Language Corpus*.

Dans ce chapitre, nous continuons la translation du SLC vers la base TILE, commencée dans le chapitre précédent avec notre travail sur les tâches de lecture.

Dans la section 12.2, nous proposons d'examiner la même tâche de détection du score à la KSS que dans le chapitre précédent, mais en utilisant la base TILE, qui contient les enregistrements de patients atteints d'hypersomnie, pour lesquels la KSS n'est pas exactement le même que dans le SLC (cf. chapitre 8). Puis, dans la section 12.3, nous irons de la tâche de détection de somnolence subjective vers une tâche de détection de la somnolence objective au long cours, mesurée par la latence moyenne à un TILE.

12.2 Détection de la somnolence instantanée sur la base TILE

12.2.1 Corpus

Nous considérons dans cette section le corpus TILE-106 (cf. chapitre 7) dont les caractéristiques de somnolence subjective à court terme sont présentées dans le tableau 12.1.

Donnée	**Femmes**	**Hommes**	**Total**
Nombre de sujets	63	43	106
Nombre d'échantillons	315	215	530
Âge (é-t)	33.9 (11.5)	38.7 (16.9)	35.9 (14.1)
Niveau social (é-t)	6.0 (2.5)	4.6 (2.3)	5.4 (2.5)
KSS (é-t)	4.6 (1.3)	4.3 (1.2)	4.4 (1.3)
Nombre d'échantillons S	72	36	108
Durée totale S	1 h 36 m 4 s	49 m 57 s	2 h 26 m
Nombre d'échantillons NS	243	179	422
Durée totale NS	4 h 58 m 22 s	4 h 31 s	8 h 58 m 53 s

TABLEAU 12.1 – Statistiques du corpus TILE. *S* : somnolent (KSS $\geq$ 7.0) ; *NS* : non-somnolent (KSS $<$ 7.0).

12.2.2 Méthode

La procédure de classification, basée sur un SVM, est strictement identique à celle proposée dans la section 11.5 du chapitre 11 précédent.

Le SLC est déjà divisé en sous-corpus d'entraînement, de développement et de test, mais ce n'est pas le cas pour la base TILE. Nous utilisons donc une validation croisée qui exclut à chaque itération un locuteur (LOSOCV) qui servira de test, puis les locuteurs restants sont divisés en bases d'entraînement et de développement (respectivement quatre cinquièmes et un cinquième des locuteurs restants), équilibrées en termes de sexe, d'âge et de KSS moyennes.

De plus, les performances sur le corpus TILE sont mesurées durant cette phase avec le score F1 (cf. chapitre 16) en raison de la validation croisée qui laisse trop peu d'échantillons dans la base de développement pour que l'UAR, utilisé pour calculer les performances dans la suite, soit pertinent.

12.2.3 Résultats

Les performances de classification du système proposé sont rapportées dans le tableau 12.2. L'application du schéma de classification sur la base TILE conduit à des performances inférieures à 50% (48.7% d'UAR). Contrairement à ce qui a été observé sur le SLC, le centrage des marqueurs vocaux semble diminuer les performances : le même schéma de classification sans centrage permet d'augmenter les performances de presque 7% d'UAR absolu (système (l), 55.6% d'UAR).

Une des raisons qui conduisent à ces faibles scores pourrait être le très fort déséquilibre entre les classes S et NS (moins de 20% des échantillons sont étiquetés S). En abaissant la limite entre les deux classes à « 5 - Ni éveillé, ni somnolent », la répartition entre les classes devient quasiment équilibrée : 251 échantillons sont étiquetés S (47.3%) et 279 NS (52.7%). Ce changement ne suffit pas cependant pas à améliorer radicalement les précédents résultats, conduisant à un UAR de 56.3% pour le système (m), soit à peine 0.7% de plus que le système (l).

	Centrage	Corpus	limKSS	UAR
(k)	oui	base TILE	7.0	48.7%
(l)	non	base TILE	7.0	55.6%
(m)	oui	base TILE	5.0	56.3%

Tableau 12.2 – Paramètres et performances des systèmes basés sur un SVM dans la base TILE.

12.2.4 Discussion

Une première explication à la différence de performances sur la détection de la somnolence subjective à court terme entre les tâches de lecture du SLC et la base TILE – à système et marqueurs identiques – pourrait être la différence de taille entre les corpus, le SLC ayant plus d'échantillons (791 au total) que la base TILE (530 au total).

Mais ces différences peuvent s'expliquer autrement que par la différence du nombre d'échantillons entre les deux corpus. En effet, au-delà de différences dans les tâches vocales ou de nombre de sujets, une différence majeure entre les corpus SLC et TILE est la façon dont la somnolence subjective à court terme est opérationnalisée.

Dans la base TILE, il s'agit d'une KSS remplie par le patient avant chaque sieste du TILE. Le score utilisé pour annoter les données est donc un reflet de l'évaluation subjective de sa somnolence par le patient lui-même, et le but du système est de retrouver, à partir de marqueurs vocaux, cette annotation subjective. L'hypothèse sous-jacente est donc que la façon dont le patient perçoit sa somnolence est identifiable à partir de sa voix, c.-à-d. qu'il existe un lien direct et spécifique entre somnolence subjective et activité vocale et/ou linguistique.

Dans la base SLC, l'annotation est mixte, moyenne d'une KSS remplie par le sujet lui-même (mesurant donc, comme précédemment, la perception que celui-ci a de sa somnolence) et de deux KSS remplies par des annotateurs externes. Il est précisé dans les articles présentant la méthodologie du SLC que ces derniers sont entraînés à annoter la somnolence, sans plus de détails méthodologiques. Le fait que le score utilisé soit la moyenne du score de deux annotateurs et d'un seul questionnaire du sujet tend à en faire une mesure principalement comportementale (jugement extérieur de manifestations comportementales associées à de la somnolence) pondérée par le ressenti subjectif du patient. Parmi ces manifestations comportementales se trouve la voix : les échantillons audio sont annotés avec un score reflétant le ressenti des investigateurs, construit lui-même en partie sur la voix du patient.

Alors qu'avec la base TILE la tâche de classification tend à identifier l'influence du ressenti subjectif de somnolence du sujet sur sa voix, la tâche induite par l'annotation du SLC est liée à la détection des manifestations comportementales de la somnolence. La seconde tâche semble ainsi moins complexe : il s'agit de retrouver, grâce à des marqueurs vocaux, un score en partie élaboré sur la voix des sujets. Cette différence de complexité de tâche pourrait expliquer les différences de performances observées entre le SLC et la base TILE.

12.3 Détection de la somnolence objective au long cours

12.3.1 Hypothèse et contexte

Cette section introduit une nouvelle tâche liée à la détection de la somnolence dans la voix : la détection de la somnolence objective au long cours. L'objectif ici n'est plus de détecter le ressenti subjectif du patient ou de retrouver le score d'un annotateur extérieur notant l'impact de la somnolence sur des éléments de comportement du sujet, mais de retrouver, grâce à la voix, le résultat de mesures polysomnographiques de l'hypersomnolence.

Les latences individuelles au TILE n'étant pas une mesure validée de la somnolence dite « objective », nous proposons un changement de paradigme, pour passer de la détection de la somnolence au court terme à la somnolence au long cours. Alors que la tâche présentée dans le chapitre précédent a pour but l'estimation d'un état ayant une durée de l'ordre de la minute, la détection de la somnolence objective au long cours a pour but l'estimation de la somnolence à plus long terme, sur des échelles de temps allant de la semaine à plusieurs mois.

Le suivi médical des patients souffrant de Somnolence Diurne Excessive peut tirer bénéfice de la détection de la somnolence à court terme, mais aussi à long terme, pour permettre aux médecins de suivre sur de longues plages de temps les variations des marqueurs de traits de somnolence des locuteurs. La détection d'une telle somnolence s'appuie sur le fait que dans le corpus TILE, chaque locuteur est enregistré cinq fois, à des moments différents de la journée. Cette partie a donc pour objectif de classer non plus les échantillons indépendamment les uns des autres, mais les locuteurs entre eux grâce aux enregistrements de leurs cinq siestes.

Dans cette section, la vérité terrain utilisée pour détecter la somnolence est la latence d'endormissement moyenne au TILE, pour laquelle une valeur inférieure à 8 minutes est considérée comme pathologique. La limite de 8 minutes sur la moyenne des latences d'endormissement est une limite médicale utilisée dans le diagnostic de nombreuses maladies telles que la narcolepsie (Aldrich *et coll.*, 1997), qui est une valeur de référence dans le milieu de la médecine du sommeil (Arand *et coll.*, 2005).

Deux méthodes sont proposées pour cette tâche. Dans la section 12.3.3, nous proposons d'utiliser la moyenne des marqueurs par locuteur comme vecteur d'entrée à un système de classification, tandis que dans la section 12.3.4, nous proposons d'estimer dans un premier temps les latences individuelles, puis de fusionner les estimations ainsi obtenues. Tous les résultats de classification de cette section sont présentés dans le tableau 12.4.

12.3.2 Base de données

La base de données utilisée dans cette section est la base TILE-106 (cf. chapitre 7). Les statistiques de ce corpus utiles à cette section sont présentées dans le tableau 12.3.

Donnée	Femmes	Hommes	Total
Nombre de sujets	63	43	106
Nombre d'échantillons	315	215	530
TILE moyen (écart-type) en minutes	11.8 (4.6)	10.4 (5.1)	11.2 (4.8)
Nombre de sujets S (TILE)	13	15	28
Durée totale d'enregistrement S (TILE)	1 h 20 m 55 s	1 h 39 m 37 s	3 h 32 s
Nombre de sujets NS (TILE)	50	28	78
Durée totale d'enregistrement NS (TILE)	5 h 13 m 30 s	3 h 10 m 52 s	8 h 24 m 22 s

Tableau 12.3 – Statistiques du corpus TILE. S : somnolent ; NS : non-somnolent.

12.3.3 Méthode n°1 : Moyenne des marqueurs par locuteur

Sélection des marqueurs avec la corrélation de Spearman

Une première intuition pour estimer la classe de somnolence des locuteurs est de faire la moyenne des cinq jeux de marqueurs de chaque locuteur (un par sieste), et d'entraîner un seul classifieur (SVM) grâce à un unique ensemble de marqueurs moyens par locuteur. Nous utilisons ensuite la même validation croisée (LOSOCV) et la même procédure de sélection des marqueurs grâce à la corrélation de Spearman que dans la section précédente.

Cette procédure conduit à un UAR atteignant à peine 50 % (système 2a).

Réf.	Système	Sélection des marqueurs	seuil TILE	UAR
(2a)	Moyenne des paramètres	Spearman	8	50.2 %
(2b)	Moyenne des paramètres	Mann-Whitney	8	54.8 %
(2c)	Fusion tardive	Spearman	8	45.6 %
(2d)	Fusion tardive	Mann-Whitney	8	53.6 %
(2e)	Fusion tardive	Mann-Whitney	13	**63.8 %**

Tableau 12.4 – Résultats des systèmes de classification pour la détection de la somnolence à long terme objective sur la base TILE.

Sélection des marqueurs avec un test de Mann-Whitney

Nous proposons une autre méthode très proche de la précédente, qui se base également sur la moyenne des marqueurs acoustiques et sur la procédure de classification, en changeant le critère de sélection des marqueurs d'une corrélation de Spearman par la *p-value* d'un test statistique de Mann-Whitney, entre les marqueurs de la classe S et ceux de la classe NS. Ainsi, au lieu d'être ordonnés par leur corrélation avec le label avant que ce dernier ne soit binarisé, les marqueurs sont ordonnés par leur pouvoir discriminant entre les deux classes de somnolence. Cette approche permet un gain de 4.6% d'UAR absolu, le système 2b atteignant un UAR de 54.8%.

12.3.4 Méthode n°2 : fusion tardive

Nous proposons une deuxième méthode, basée sur la fusion tardive des probabilités d'appartenance à la classe « Somnolent » sur les cinq siestes. Ainsi, nous attribuons une classe (S ou NS) à chacun des cinq échantillons vocaux d'un sujet, indépendamment les uns des autres, avec la même limite de 8 minutes utilisée pour la latence moyenne. Nous entraînons

ensuite un classifieur (SVM) à estimer les probabilités d'appartenance à la classe S et à la classe NS des échantillons (notés respectivement p_i et $\overline{p_i}$ pour le $i$ème échantillon de chaque locuteur).

En moyennant les probabilités inférées des cinq échantillons d'un même locuteur, nous obtenons ainsi sa probabilité moyenne d'appartenir à chacune des classes p_{moy} (et $\overline{p_{moy}}$ pour les NS). En prenant ensuite le maximum entre p_{moy} et $\overline{p_{moy}}$, la classe du locuteur est déterminée. Cette procédure est résumée dans la figure 12.1.

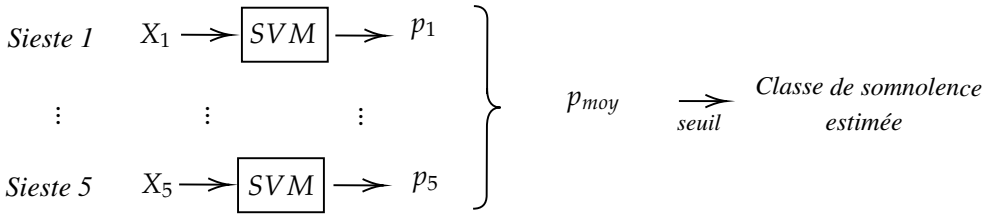

FIGURE 12.1 – Procédure pour l'estimation de la classe de somnolence selon la méthode n°2. X_i : ensemble des paramètres vocaux triés pour la sieste i. p_i : probabilité que l'échantillon de la sieste i provienne d'un locuteur somnolent. Les classifieurs SVM représentés sont tous identiques et ont été entraînés sur tous les échantillons, toutes itérations confondues.

Les deux systèmes basés sur la fusion tardive (2c) et (2d), sélectionnant les marqueurs respectivement avec une régression entre marqueurs vocaux et TILE moyen, et un test de Mann-Whitney entre les deux classes S et NS, atteignent des scores inférieurs à ceux obtenus avec les systèmes précédents (respectivement 45.6% et 53.6%).

12.3.5 Discussion

Les résultats de classification obtenus sont trop faibles pour une utilisation en situation réelle, qui nécessiterait au moins 80% de performances. Nous faisons cinq hypothèses pour expliquer ces résultats.

Première hypothèse : choix de la limite entre les classes de somnolence Une première hypothèse pour expliquer l'échec de cette approche pourrait être la limite de 8 minutes, choisie pour délimiter les deux classes de somnolence sur la base des recommandations de l'*American Academy of Sleep Medicine* (Arand *et coll.*, 2005), et qui pourraient ne pas correspondre au seuil d'expression de la somnolence dans la voix.

De même que dans le précédent chapitre 11, nous retestons notre système avec une autre limite pour séparer les deux classes de somnolence selon la valeur de TILE moyenne des patients. Afin d'avoir un meilleur équilibre entre les classes, nous proposons la limite de 13 minutes, avec 63 locuteurs « somnolents » (TILE moy $\leq$ 13 min.) et 43 locuteurs « non somnolents » (TILE moy $>$ 13 min). En réappliquant les systèmes précédents avec cette nouvelle limite, nous obtenons un score de presque 64 % (2e), ce qui représente une amélioration de plus de 10 % par rapport au système précédent, mais malgré tout inférieur aux performances nécessaires à une exploitation clinique du système.

Deuxième hypothèse : influence des comorbidités Deuxièmement, de même que le stress ou les émotions peuvent influencer l'expression de la somnolence immédiate dans la voix, l'anxiété, la dépression, et une multitude d'autres facteurs propres au locuteur peuvent également polluer les marqueurs vocaux utilisés pour la détection de la somnolence. Le corpus

étant composé d'enregistrements de patients souffrant de SDE, la plupart ont des facteurs de comorbidité qui pourraient influencer leur voix, et interférer dans la reconnaissance de la somnolence dans celle-ci.

Troisième hypothèse : entraînement du classifieur Une cause de ces résultats pourrait se trouver dans la façon dont le classifieur est entraîné. Non seulement un unique classifieur est entraîné sur toutes les siestes de manière indistincte afin d'augmenter la quantité de données d'entraînement, mais le fait de moyenner les probabilités des cinq échantillons de manière égale masque également l'éventuelle importance que pourraient avoir certaines siestes par rapport à d'autres.

En effet, chaque sieste du TILE a des caractéristiques qui lui sont propres : par exemple, le repas est généralement servi peu avant la sieste de 13h, ou lors de la dernière sieste des TILE, les patients rapportent généralement des signes de fatigue et d'impatience, au point que la communauté de médecine du sommeil se questionne sur l'utilité de cette 5e sieste (Muza *et coll.*, 2016).

Un classifieur par sieste ou une pondération des différentes probabilités semblent des pistes de recherches intéressantes, qui pourraient mener à une meilleure compréhension des phénomènes mis en jeu et à de meilleures performances.

Quatrième hypothèse : sensibilité des marqueurs acoustiques Une quatrième hypothèse à ne pas négliger concerne la sensibilité des marqueurs acoustiques à la somnolence. L'hypothèse sous-jacente à ces systèmes de classification était que les marqueurs acoustiques peuvent être une mesure de la somnolence (cf. chapitre 1). En plus de leur manque de spécificité au regard des autres comorbidités, les marqueurs acoustiques ne sont peut-être pas suffisamment sensibles à eux seuls pour permettre la détection de la somnolence dans la voix.

Cinquième hypothèse : estimation automatique des paramètres Enfin, une cinquième hypothèse est que lorsque les sujets sont somnolents, les systèmes d'estimation automatique des paramètres n'ont plus les mêmes performances que sur de la parole de sujet sain, et que les paramètres renvoyés par ces systèmes ne reflètent plus la réelle nature du signal de parole contenu dans les enregistrements audio.

12.4 Conclusion et perspectives

En conclusion, le système basé sur des marqueurs acoustiques établis dans le chapitre 11 précédent ne permet pas d'atteindre des performances de classification de la somnolence objective au long cours suffisantes pour espérer une application clinique de celui-ci.

Deux perspectives de recherche nous semblent prometteuses pour remédier à cette situation. D'une part, un système prenant en compte les spécificités de chaque sieste (comme par exemple utiliser un classifieur par sieste) permettrait d'affiner la façon dont sont ensuite fusionnées les probabilités pour déterminer le statut du locuteur (S ou NS). Cette piste de recherche est laissée en suspens et n'est pas traitée dans la suite de ce document.

Par ailleurs, nous souhaitons approfondir la quatrième hypothèse exposée ci-dessus, en cherchant de nouveaux marqueurs plus sensibles à la somnolence que ceux que nous avons conçus. L'idée d'utiliser les marqueurs dérivés de systèmes d'apprentissage profond, comme cela a été proposé pour le challenge IS19 (cf. chapitre 11) nous est proscrite, car elle ne remplit pas notre impératif d'explicabilité. Nous proposons ainsi dans les prochains chapitres de

nouveaux marqueurs vocaux de la somnolence, qui ont pour but d'évaluer l'impact de celle-ci sur les capacités cognitives du lecteur.

Chapitre 13

Erreurs de lecture

Le travail décrit dans ce chapitre (définition des erreurs de lecture et annotations) a été réalisé en collaboration avec Mathilde Rieant (MR) et Gabrielle Chapouthier (GC), accueillies au LaBRI dans le cadre de leur stage de M1 pour le cursus orthophonie du CFUOB.

13.1 Objectifs et précédents travaux

13.1.1 Objectifs

Dans ce chapitre, nous proposons une nouvelle approche pour la détection de la somnolence au long cours, en utilisant les erreurs effectuées lors de la lecture de textes à voix haute.

13.1.2 Précédents travaux

Concurremment à nos travaux, une autre étude, portant sur la prédiction des déficiences cognitives chez des patients atteints de la maladie de Parkinson, s'est intéressée aux erreurs de lectures, à la fois annotées manuellement et automatiquement (Romana *et coll.*, 2021).

Grâce aux annotations manuelles des erreurs couramment utilisées en transcription automatique de la parole (insertions, délétions, substitutions), le système proposé a obtenu un score de 0.64 de Coefficient de corrélation de concordance (CCC) entre le score estimé et le score réel, obtenant des performances plus élevées que le même système basé sur des marqueurs acoustiques (0.33 de CCC).

Malgré la différence de tâche entre cette étude et notre étude, cela nous conforte dans notre hypothèse que les erreurs de lecture peuvent être des marqueurs pertinents de la somnolence.

13.2 Annotation des erreurs de lecture

13.2.1 Procédure

La base de données utilisée dans cette partie est la base TILE-106, décrite dans le chapitre 7. Pour chaque sujet et chaque enregistrement, les erreurs de lecture ont été annotées manuellement par MR, GC, ou moi-même, directement lors de la supervision des enregistrements de la base TILE au pôle universitaire de médecine du sommeil de Bordeaux et/ou à partir des enregistrements inclus dans la base de données. Toute incertitude a été résolue par consensus.

13.2.2 Liste des erreurs de lecture

Nous avons retenu cinq catégories d'erreurs, afin de différencier différents comportements de lecture tout en obtenant un nombre suffisant d'observations dans chaque catégorie, afin de garder une certaine généralité et de pouvoir comparer les sujets les uns par rapport aux autres sur des critères communs.

Achoppements (Ach) : « hésitation, coupure, dans le rythme de la parole » (Brin *et coll.*, 2018). Ces erreurs sont un reflet de la capacité d'assemblage du lecteur, c'est-à-dire sa capacité de mettre bout à bout des syllabes pour former un mot (cf chapitre 1). Ainsi, lorsque le lecteur commence la lecture d'un mot, s'arrête, et se reprend, le processus d'assemblage a été interrompu, causant un achoppement. Nous n'avons pas pris en compte les arrêts entre les mots, mais seulement les arrêts qui se produisent au milieu d'un mot, ou les allongements

artificiels de certaines voyelles, qui témoignent d'une hésitation. Dans le cas de la reprise d'une phrase ou d'un bout de phrase, un seul achoppement est compté, quelle que soit la longueur de la reprise.

Paralexies (Plx) : « erreur d'identification de mots écrits consistant à oraliser un mot écrit à la place d'un autre » (Brin *et coll.*, 2018). Contrairement aux achoppements, les paralexies reflètent les erreurs d'adressage du lecteur. La capacité d'adressage est le fait de lire un mot dans sa globalité, sans le découper en syllabes ou le déchiffrer, dont les paralexies sont des erreurs symptomatiques (cf chapitre 1). Nous avons généralisé cette catégorie à toute prononciation d'un mot, existant ou non, qui est lu à la place du mot correct. Les télescopages (oublis d'une ou plusieurs syllabes dans un mot) sont donc inclus dans cette catégorie.

Oublis de mots (O) : cette erreur est comptée lorsque le lecteur oublie de lire un mot et passe directement au début du mot suivant.

Additions de mots (Add) : cette erreur est comptée lorsque le lecteur ajoute un mot qui n'était pas dans le texte original.

Inversion de mots (Inv) : cette erreur est observée lorsque le locuteur inverse plusieurs mots dans la phrase. Si un locuteur se reprend après une paralexie, un oubli ou une addition, aucun achoppement supplémentaire n'est compté, sauf s'il se trompe lors de la reprise, auquel cas l'erreur de reprise est identifiée et comptabilisée.

13.3 Sensibilité des erreurs de lecture à la somnolence

Ce chapitre traitant de la détection de traits des locuteurs sur le corpus TILE-106 (cf chapitre 7), les patients seront annotés comme « Somnolent » si la valeur de leur latence moyenne d'endormissement au TILE est inférieure ou égale à 8 minutes.

Afin de mesurer si les erreurs élaborées précédemment varient avec la somnolence, les distributions du nombre total de chaque type d'erreur par locuteur chez les patients somnolents et non somnolents sont représentées dans la figure 13.1 (moyenne $\pm$ SEM – erreur standard de la moyenne). Sur tous les types d'erreurs sauf les inversions de mots, les patients somnolents font plus d'erreurs que leurs homologues non somnolents (tests de Mann-Whitney. Ach : $U = 873, p = 8.1 \times 10^{-2}$; O : $U = 738, p = 7.8 \times 10^{-3}$; Add : $U = 847, p = 5.0 \times 10^{-2}$; Plx : $U = 759, p = 1.2 \times 10^{-2}$; Inv : $U = 1041, p = 0.72$; total : $U = 765, p = 1.4 \times 10^{-2}$). Par ailleurs, ces dernières sont très peu représentées et elles ne sont pas observées sur de nombreux locuteurs. En conséquence, nous choisissons de ne pas considérer ce type d'erreur dans la suite de l'analyse.

13.4 Étude des sources d'influence de production d'erreurs

Il est nécessaire de pouvoir séparer l'influence de la somnolence des facteurs extérieurs pouvant provoquer ces erreurs. Ces facteurs peuvent être les différences entre les textes (différence de taille, quantité de dialogues, difficulté du texte) ou les différents facteurs temporels tels que la prise de repas ou la fatigue accumulée de la journée. Dans la suite, « influence de l'itération » désignera l'influence de tels facteurs sur les erreurs produites par le locuteur. Afin de séparer la contribution de la somnolence de celle de l'itération, nous avons appliqué à nos

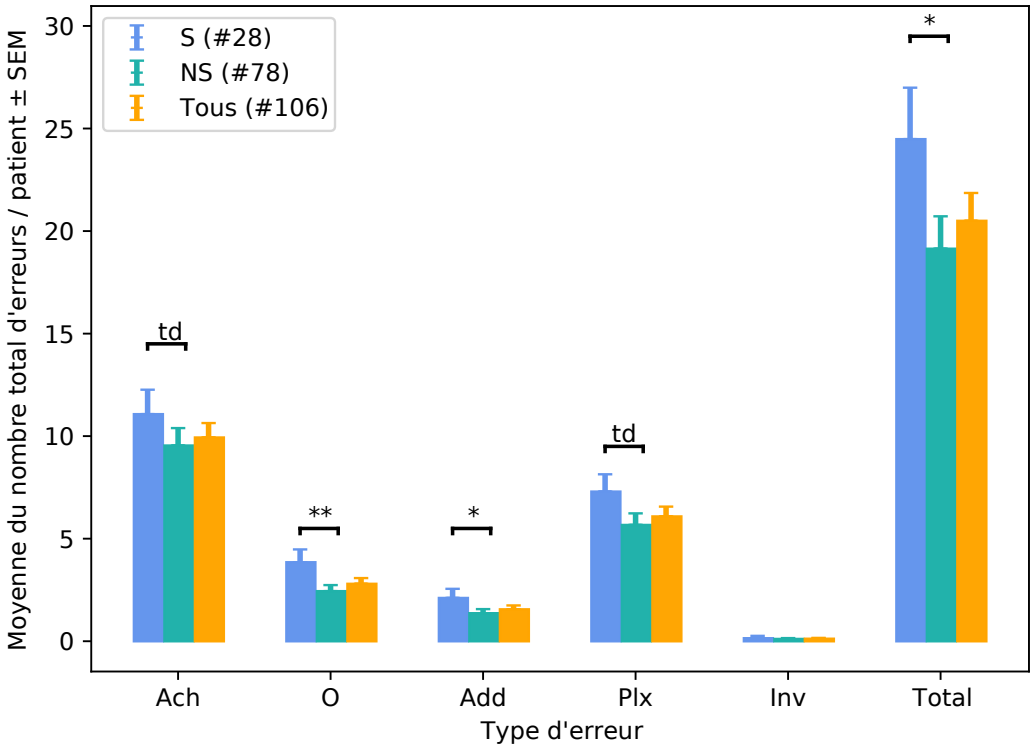

FIGURE 13.1 – Distribution du nombre total d'erreurs par locuteur (moyenne ± SEM). Ach : achoppements, O : oublis, Add : additions, Plx : paralexies, Inv : inversion de mots. Tests de Mann-Whitney (td : $p < 10^{-1}$, * : $p < 5 \times 10^{-2}$, ** : $p < 10^{-2}$).

données une ANOVA multivariée à mesures répétées prenant en compte l'itération, la latence d'endormissement au TILE et le score de la KSS.

13.4.1 Additions de mots

La somnolence objective a une influence quasiment significative sur les variations intersujets du nombre d'additions (influence de la valeur de TILE sur les variations intersujets : $F = 3.5$; $p = 6.6 \times 10^{-2}$), tandis que la somnolence subjective a un effet quasiment significatif sur les variations intersujets du nombre d'additions (influence de la KSS sur les variations intrasujets : $F = 3.8$; $p = 5.2 \times 10^{-2}$).

Cela signifie que les différences observées entre les sujets indépendamment du temps sont principalement expliquées par leurs différences de TILE (ce qui confirme le lien entre TILE et additions) tandis que celles observées sur chaque sujet au cours du temps (influences conjointes de la session et du locuteur) sont principalement expliquées par les différences de variation de KSS au cours des itérations du test.

La session n'a aucun effet significatif sur la production des additions. Nous faisons donc l'hypothèse que les variations du nombre d'additions sont principalement dues à celles des somnolences objectives et subjectives, et qu'elles sont donc indépendantes du texte et des autres effets d'itération.

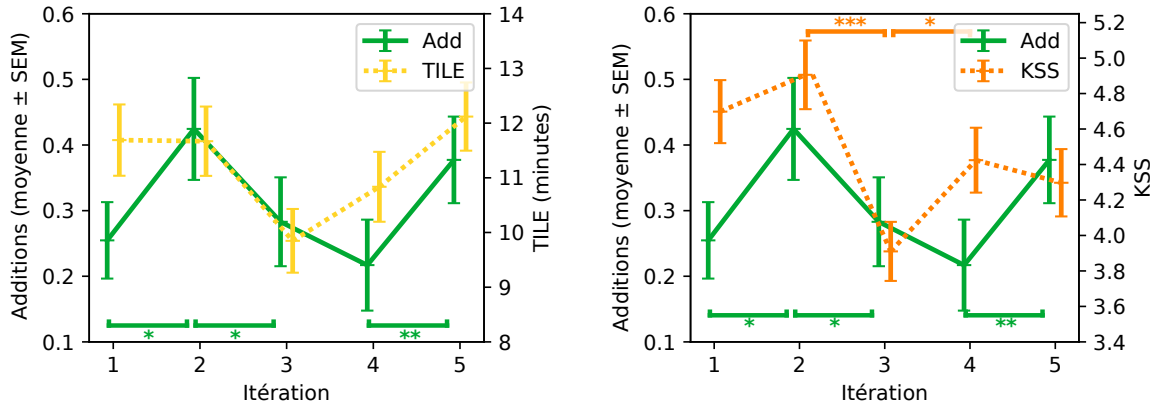

FIGURE 13.2 – Nombre d'additions comparé au TILE (gauche) et à la KSS (droite) (moyenne ± SEM). Tests de Mann-Whitney (td : $p < 1 \times 10^{-1}$, * : $p < 5 \times 10^{-2}$,** : $p < 10^{-2}$, *** : $p < 10^{-3}$).

Les variations du nombre d'additions au regard des variations du TILE et de la KSS en fonction des itérations du TILE sont représentées dans la figure 13.2.

13.4.2 Oublis de mots

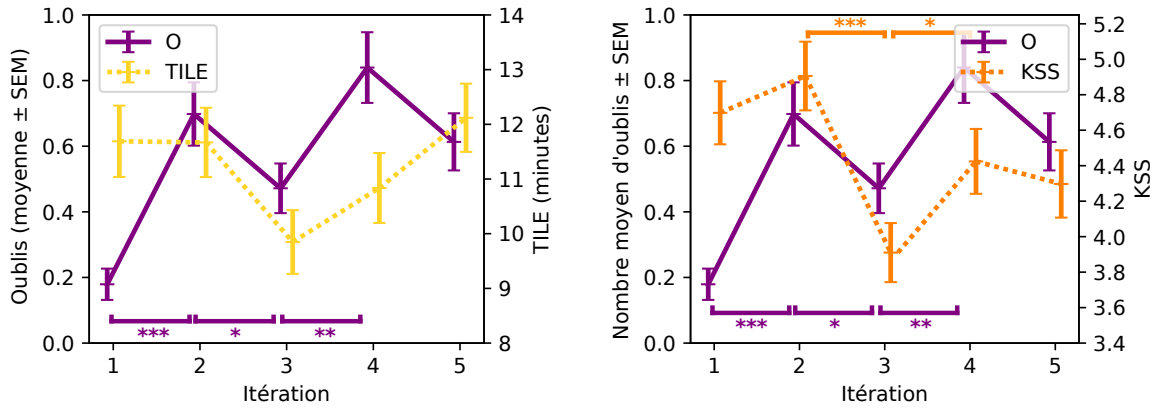

FIGURE 13.3 – Nombre d'oublis comparé au TILE (gauche) et à la KSS (droite) (moyenne ± SEM). Tests de Mann-Whitney (td : $p < 1 \times 10^{-1}$, * : $p < 5 \times 10^{-2}$,** : $p < 10^{-2}$, *** : $p < 10^{-3}$).

De même, la somnolence objective a une influence sur le nombre d'oublis de mots entre les locuteurs (influence de la valeur de TILE sur les variations intersujets : $F = 3.2$; $p = 7.5 \times 10^{-2}$) tandis que la somnolence subjective a une influence sur les variations du nombre d'oublis de mots au cours des sessions (influence de la KSS sur les variations intrasujets : $F = 3.1$; $p = 8.1 \times 10^{-2}$). En revanche, contrairement aux additions, ces erreurs subissent également les effets de l'itération (effet de l'itération sur les variations intrasujets : $F = 12.0$; $p = 3.0 \times 10^{-9}$).

Les variations du nombre d'oublis au regard des variations du TILE et de la KSS en fonction des itérations du TILE sont représentées dans la figure 13.3.

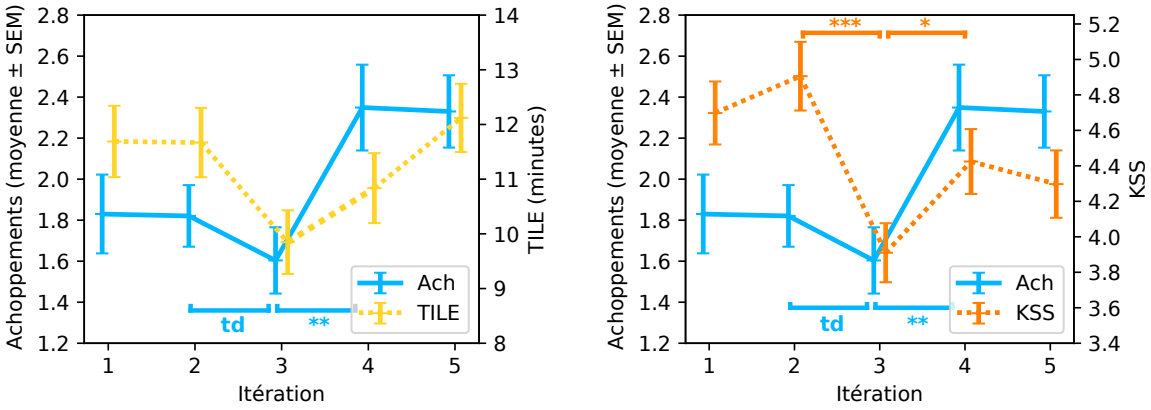

FIGURE 13.4 – Nombre d'achoppements comparé au TILE (gauche) et à la KSS (droite) (moyenne ± SEM). Tests de Mann-Whitney (td : $p < 1 \times 10^{-1}$, * : $p < 5 \times 10^{-2}$,** : $p < 10^{-2}$, *** : $p < 10^{-3}$).

13.4.3 Achoppements

De même que pour les oublis de mots, l'étude des divers effets ayant une influence sur les variations du nombre d'achoppements permet de mettre en évidence une influence significative de la KSS ($F = 4.2$; $p = 4.2 \times 10^{-2}$) et de l'itération ($F = 7.4$; $p = 9.4 \times 10^{-6}$) sur les variations intrasujets de ce type d'erreur. En revanche, la valeur de TILE ne semble avoir aucune influence sur ce type d'erreur, et aucun effet significatif n'a pu être trouvé pour expliquer les variations intersujets.

13.4.4 Paralexies

Les variations de paralexies au cours des itérations ne sont influencées que par les facteurs d'itération (effet de l'itération sur les variations intrasujets : $F = 6.0$; $p = 1.1 \times 10^{-4}$).

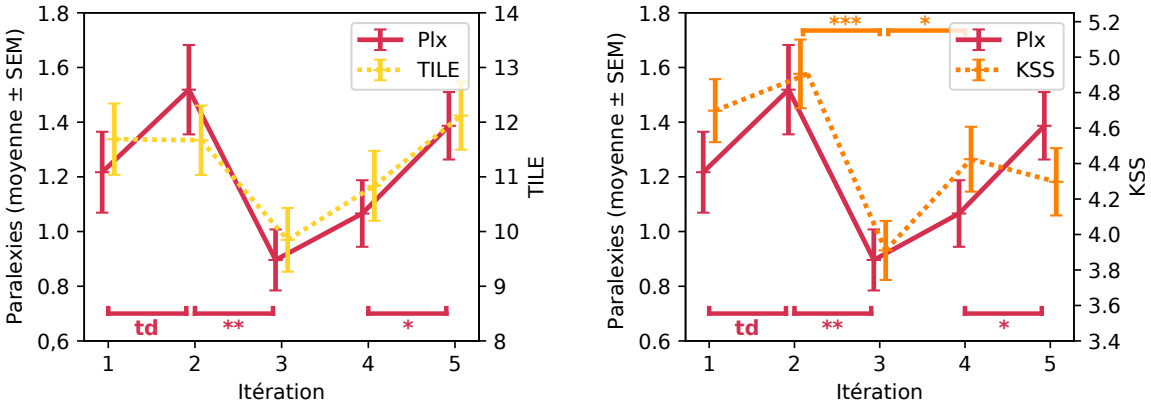

FIGURE 13.5 – Nombre de paralexies comparé au TILE (gauche) et à la KSS (droite) (moyenne ± SEM). Tests de Mann-Whitney (td : $p < 1 \times 10^{-1}$, * : $p < 5 \times 10^{-2}$,** : $p < 10^{-2}$, *** : $p < 10^{-3}$).

13.5 Estimation de la somnolence du locuteur

13.5.1 Méthode

Nous utilisons les erreurs de lecture précédemment élaborées ayant comme facteur d'influence la somnolence comme marqueurs pour un système de classification d'état du locuteur. Pour cela, nous concaténons les erreurs de lecture des cinq siestes pour chaque locuteur (fusion précoce), dont nous nous servons comme entrée à un classifieur (SVM avec un noyau linéaire).

De la même façon que dans le chapitre précédent, la classification est effectuée en utilisant une validation croisée du type *Leave One Speaker Out Cross Validation* (LOSOCV).

13.5.2 Résultats

La matrice de confusion correspondante est représentée dans le tableau 13.1 (a) dont l'UAR atteint 78.7 %.

Par ailleurs, les paralexies n'étant pas liées à la somnolence, mais uniquement à des effets d'itération, la même procédure sans considérer ce type d'erreur conduit à la matrice de confusion présentée dans le tableau 13.1 (b). L'UAR associé est de 82.6 %.

(a)	S_{pred}	NS_{pred}
S_{th}	23	4
NS_{th}	22	57

UAR = 78.7 %

(b)	S_{pred}	NS_{pred}
S_{th}	20	7
NS_{th}	7	72

UAR = 82.6 %

TABLEAU 13.1 – Matrices de confusion et UAR des classifieurs utilisant les erreurs de lecture comme marqueurs de la somnolence, pour la détection d'une latence moyenne d'endormissement au TILE pathologique ($\leq$ 8 minutes). (a) En prenant en compte toutes les erreurs de lecture. (b) En excluant les paralexies.

13.6 Analyse des marqueurs sélectionnés

Les erreurs de lecture semblent de bons marqueurs de l'état de somnolence objective des locuteurs. Cependant, en raison de leur définition ou du texte incitant ou non ces erreurs, elles n'ont pas toutes la même importance dans la détection de la somnolence. En effet, en moyennant les coefficients attribués par le SVM aux différentes erreurs selon les différentes itérations de la validation croisée, nous obtenons les quatre marqueurs suivants les plus importants dans la classification : les additions des premières et cinquièmes siestes (ayant des coefficients respectifs $c = 8.0 \times 10^{-2}$ et $c = -1.5 \times 10^{-1}$), les achoppements de la troisième sieste ($c = 9.7 \times 10^{-2}$) et les oublis de la quatrième sieste ($c = -1.3 \times 10^{-1}$).

Ces coefficients sont cohérents avec les résultats de la section précédente. En effet, les additions, qui ont ici le plus de poids dans la prise de décision du niveau de somnolence, avaient été identifiées comme ne dépendant que de la somnolence objective concernant les variations intersujets et ne dépendant pas des effets d'itération. De même, le deuxième marqueur le plus important dans la prise de décision est les oublis, qui malgré les effets d'itération varient avec la somnolence objective. Enfin, même si l'étude statistique des achoppements n'avait pas mis en valeur d'influence de la valeur de TILE sur la production de ce type d'erreur, leur contribution dans la prise de décision n'est pas négligeable.

Erreur	Session	Coef.	ANOVA
Additions	1	8.0×10^{-2}	KSS : td
	2	-1.5×10^{-1}	TILE : td
Achoppements	3	9.7×10^{-2}	TILE : td
Oublis	4	-1.3×10^{-1}	TILE : td

TABLEAU 13.2 – Quatre marqueurs les plus importants dans la prise de décision par le classifieur SVM. Coef. : Coefficient associé au marqueur, ANOVA : facteurs significatifs dans les différences intersujets lors du calcul de l'ANOVA de la section précédente.

13.7 Discussion

La répartition des coefficients de manière inégale sur les différentes siestes du test pose la question de l'importance relative des itérations pour l'estimation du niveau global de somnolence des locuteurs. En effet, si les additions sont les marqueurs ayant le plus de poids sur la première et la dernière sieste, leur contribution pour la détection de l'état du locuteur lors de la troisième sieste est très faible ($c = 8.8 \times 10^{-3}$), alors que celle des achoppements est la plus importante. Une cause probable de ces disparités est l'inégalité de contenu des textes. En effet, de nombreuses erreurs du corpus se répètent et certains mots sont systématiquement la cible d'une erreur spécifique. Par exemple, « méditatif » est très souvent prononcé « médiatif », causant de nombreuses paralexies à la cinquième sieste, ou encore « Il me répéta alors » est souvent lu à la place de « Et il me répéta alors », causant de nombreux oublis à la troisième sieste. Cela souligne l'aspect capital du choix des textes lus pour l'utilisation des erreurs de lecture en tant que marqueurs de la somnolence.

Par ailleurs, la définition des erreurs a également une influence sur leur robustesse. Nous faisons effectivement l'hypothèse que notre définition des achoppements ne prenant en compte que les interruptions au sein des mots et non entre les mots induit un biais qui empêche le marqueur de refléter l'état de somnolence du locuteur. De même, la fusion des paralexies et des télescopages dans la même catégorie pourrait induire des biais qui réduisent leur intérêt comme marqueurs de la somnolence.

13.8 Conclusion et perspectives

En conclusion, nous avons proposé un nouveau type de marqueurs pour la détection de la somnolence objective au long cours dans la voix, dont l'efficacité d'utilisation dans un système de classification permettrait leur implémentation dans des conditions cliniques réelles.

Cependant, cette étude reste à l'état de preuve de concept : cette technique n'est pas implémentable en conditions réelles en raison de son coût à la fois en termes de temps (chaque annotation requiert l'écoute de 3 à 5 fois de chaque enregistrement) et de formation (l'annotation des erreurs telles qu'elles sont définies dans ce chapitre nécessite un entraînement spécifique). De plus, les annotations telles qu'elles sont définies précédemment ne constituent pas une vérité terrain fiable et reproductible : dans de nombreux cas, la différence entre une paralexie et une addition+deletion n'était pas claire et a demandé un consensus entre annotateurs. Cet aspect n'a pas été étudié dans les travaux présentés dans ce manuscrit, notamment en raison du coût humain de l'annotation qui n'a pas pu être faite en multiple aveugle, mais nécessiterait d'être approfondi dans des travaux futurs.

Chapitre 14

Erreurs de systèmes de transcription automatique de la parole

Sommaire

14.1 Objectifs et précédents travaux

14.1.1 Objectifs

Nous avons vu dans le chapitre précédent que les erreurs de lecture sont de bons marqueurs de la somnolence au long cours, mais ont comme désavantage notable de nécessiter une annotation manuelle. Afin d'automatiser leur extraction, nous nous appuyons donc sur des systèmes de transcription automatique (STA), qui retranscrivent de manière automatique le contenu des échantillons audio de notre base de données.

Les marqueurs sont présentés dans la section 14.2 et une première analyse statistique est proposée dans la section 14.3. Nous proposons ensuite d'utiliser ces marqueurs en combinaison avec les précédents pour détecter la somnolence au long cours subjective 14.4 et objective 14.5.

Ce chapitre utilise les systèmes de transcription automatique conçus et implémentés par Florian Boyer dans le cadre de sa thèse de doctorat, soutenue en 2021 (Boyer, 2021).

14.1.2 Corpus

Le corpus utilisé dans cette partie est la base TILE-93 (cf. chapitre 7 et annexe E).

14.1.3 Précédents travaux

L'étude sur la maladie de Parkinson (Romana *et coll.*, 2021) mentionnée dans le chapitre précédent a étudié non seulement les erreurs de lecture annotées manuellement, mais aussi grâce à un système de transcription automatique. En comparant les insertions, délétions et substitutions de mots faites par le système `Mozilla DeepSpeech` (Hannun *et coll.*, 2014) avec un entraînement affiné sur la tâche étudiée et des annotations manuelles, cette étude a étudié les enjeux de l'automatisation de l'extraction de tels marqueurs.

Ainsi, le système entièrement automatisé obtient des performances plus faibles que celui reposant sur des annotations manuelles (0.64 de coefficient de corrélation de concordance entre le score estimé et le score réel pour les annotations manuelles, contre 0.47 pour le système entièrement automatisé), mais ce dernier obtient malgré tout des performances plus élevées que le même système basé sur des marqueurs acoustiques (0.33 de CCC).

Encore une fois, malgré la différence de tâche entre cette étude et la nôtre, cela nous conforte dans notre hypothèse que les erreurs de STA peuvent être des marqueurs pertinents de la somnolence.

14.2 Description des systèmes et des marqueurs

14.2.1 Systèmes de transcription

Les systèmes de transcriptions utilisés ici sont des systèmes bout-en-bout, constitués d'un modèle acoustique et éventuellement d'un modèle de langage. Alors que le premier a pour but d'extraire une information lexicale à partir du signal acoustique de la voix, le second, basé uniquement sur du texte, a pour but de contraindre le système à l'aide de règles de grammaire qui lient les unités décodées. Les modèles de langage utilisés ont tous été entraînés sur le corpus ESTER (Galliano *et coll.*, 2009), auquel nous avons rajouté pour certains les textes cibles lus par les patients (entraînement affiné).

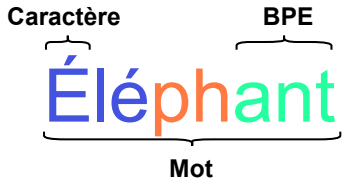

FIGURE 14.1 – Exemples des différentes unités utilisées par les systèmes de reconnaissance automatique.

Ces unités peuvent être soit des mots (systèmes traditionnels), soit plus récemment des caractères ou des portions de mots (cf figure 14.1). Dans ce dernier cas, la délimitation de ce qu'est une portion de mot n'est pas supervisée, et est gérée par le système bout-en-bout (*Byte-Pair Encoding* – BPE). Pour plus de détails, nous redirigeons le lecteur vers la thèse de Florian Boyer (Boyer, 2021).

Dans cette étude, nous nous limitons aux systèmes basés sur des caractères ou des portions de mots en raison de leur aspect novateur. Les unités des modèles de langages ont été choisies parmi les modèles disponibles au moment de l'écriture des articles que nous avons publiés sur ce sujet. Ces systèmes ont été conservés dans la suite afin d'assurer une comparaison valide avec les résultats présentés dans ce chapitre.

Par ailleurs, nous nous intéressons uniquement aux erreurs faites au niveau des portions de mot ou des mots inférés, afin de pouvoir vérifier facilement les différences entre l'hypothèse des systèmes et le texte original.

Les sept systèmes étudiés dans ce chapitre sont présentés dans le tableau 14.1.

Réf.	Unité acoustique	Unité du modèle de langage	Entraînement affiné
1	Caractères	-	-
2		Caractères	Non
3			Oui
4		Mots	Non
5			Oui
6	BPE	-	-
7		BPE	Non

TABLEAU 14.1 – Systèmes de transcription automatique considérés dans ce manuscrit. BPE : *Byte-pair encoding*.

14.2.2 Erreurs calculées

Définition

La définition des erreurs de lecture proposées dans le chapitre précédent étant complexe, nous nous concentrons dans un premier temps sur les quatre erreurs de transcription telles qu'elles sont habituellement définies pour évaluer les systèmes de transcription automatique : les insertions (Ins), les substitutions (Sub), les délétions (Del) et les unités correctement détectées (Correct).

Au décompte du nombre de chacun de ces types d'erreurs dans les échantillons, nous rajoutons le ratio de chacun des marqueurs précédent par rapport au nombre total d'unités détectées par le système. Dans ces travaux, nous nous limitons aux métriques calculées sur les mots ou sur les unités du modèle de langage (caractères ou BPE). Cela conduit à un ensemble

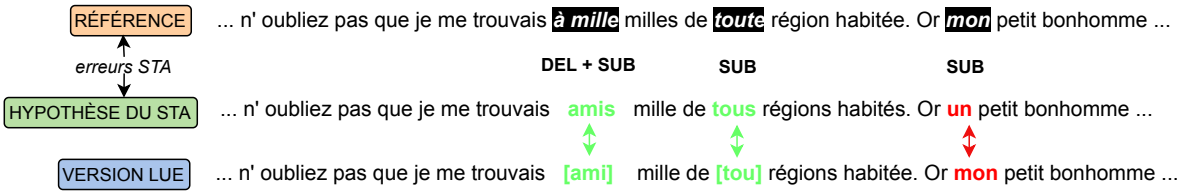

FIGURE 14.2 – Exemple de calcul d'erreurs des systèmes de transcription, pour le texte n°2 proposé lors des TILE. Les erreurs colorées en vert correspondent à des différences entre la version lue et l'hypothèse de STA, tandis que l'erreur en rouge correspond à une erreur intrinsèque aux STA, qui ne sont pas parfaits.

de 112 marqueurs d'erreurs des STA (4 erreurs calculées selon 2 modalités, en nombre et en ratio, sur 7 systèmes différents).

Exemple

Un exemple d'erreurs de STA est présenté dans la figure 14.2.

14.2.3 Différence avec les erreurs de lecture

Nous calculons ces erreurs en comparant le texte de référence et la sortie du système de transcription. Ainsi, deux sources d'erreurs sont prises en compte dans les marqueurs étudiés :

- Les erreurs de lecture, qui correspondent à une différence entre le texte de référence et ce qui est dit par le lecteur (en vert sur la figure 14.2). Dans le cas où le système de transcription était parfait (0% de taux d'erreurs de mots), ces erreurs de transcription correspondraient, à quelques transformations près, aux erreurs de lecture définies précédemment. Deux différences subsisteraient malgré tout entre les erreurs des STA et les erreurs de lecture :
  - d'une part, les reprises ne sont pas prises en compte par les STA : lors de la lecture, les patients peuvent s'arrêter au milieu d'un mot pour reprendre une portion de phrase dont ils ont analysé a posteriori qu'ils ne l'ont pas lue correctement. Lors de l'annotation manuelle, cela est aisément pris en compte avec les achoppements, au contraire du système de transcription qui comprendra la portion de mot lue avant le retour en arrière comme un mot à part entière ;
  - d'autre part, la transformation des erreurs du système de transcription en paralexies n'est pas triviale. En effet, la différence entre une paralexie et une addition+deletion ne tient qu'à la proximité, évaluée par l'annotateur, entre le mot lu et le mot attendu. Cette correspondance pourrait éventuellement être proposée en travaillant sur des transcriptions phonémiques, mais, comme précisé dans le paragraphe précédent, nous nous concentrons ici uniquement sur des systèmes proposant des transcriptions orthographiques.
- Les erreurs des systèmes de transcription, entre ce qui est dit par le lecteur et la transcription en texte par les STA, qui ne sont pas parfaits et qui font des erreurs intrinsèques dues aux architectures et aux bases de données qui ont servi à leur entraînement (en rouge sur la figure 14.2). En revanche, nous supposons que ces erreurs peuvent elles aussi nous informer sur l'état des patients : lorsque ces derniers sont somnolents, la qualité articulatoire de leur diction diminue, leur prosodie est altérée, mettant en difficulté les systèmes de transcription, qui font alors plus d'erreurs.

14.3 Première analyse statistique

Une première approche statistique de l'utilisation des erreurs des systèmes de transcription a été proposée sur la base TILE-93 dans un poster présenté aux 9e Journées de Phonétique Clinique, en mai 2021 (Martin *et coll.*, 2021).

14.3.1 Tâches investiguées

Dans cette section, trois tâches sont investiguées : la détection de la propension à l'endormissement diurne dans des conditions favorables à l'endormissement (latence moyenne d'endormissement au TILE inférieure à 8 minutes) ; la détection de la somnolence diurne excessive chez des patients hypersomniaques (ESS supérieure à 10) ; et enfin la somnolence moyenne sur une journée (moyenne des KSS lors d'un TILE supérieur à 5). Ces tâches sont décrites précisément dans le chapitre 17. Un rapide aperçu de la répartition des patients dans les catégories Somnolent et Non-Somnolent suivant les dimensions considérées et le seuil choisi est proposé dans le tableau 14.2.

Mesure	Limite S	S	NS
TILE moy.	$\leq$ 8 min.	21	72
ESS	$>$ 10	79	14
KSS moy.	$>$ 5	27	66

Tableau 14.2 – Répartition des patients dans les classes de somnolence en fonction des tâches.

14.3.2 Méthode

Pour chacun des 7 systèmes décrits dans la section précédente, nous avons calculé la moyenne de chaque type d'erreur au cours des 5 itérations du TILE (N = 112 marqueurs). Puis, nous sélectionnons les potentiels biomarqueurs de la somnolence en deux étapes.

Spécificité à la somnolence Dans le chapitre 12, nous soulevions dans les hypothèses expliquant l'échec de la classification de la somnolence uniquement à partir de marqueurs acoustiques la problématique de la spécificité des marqueurs à la somnolence, au regard d'autres caractéristiques des patients (âge, sexe, IMC ...) qui pourraient aussi influencer la voix des sujets. Afin de prendre en compte cette interférence, nous proposons dans ce paragraphe d'éliminer les marqueurs qui corrèlent avec un facteur confondant. Pour cela, nous calculons la corrélation (ρ de Spearman) des erreurs des STA avec différentes caractéristiques des locuteurs qui ont été collectées dans la base TILE et que nous avons identifiées comme pouvant potentiellement être des sources d'erreurs pour le système de transcription. Ces caractéristiques sont :

- le sexe ;
- l'âge ;
- l'Indice de Masse Corporelle (IMC) ;
- le tour de cou ;
- le niveau sociodémographique ;
- le niveau d'anxiété ;
- le niveau de dépression ;
- la pathologie.

La corrélation entre le sexe et les erreurs des STA est mesurée grâce à un test de Mann-Whitney entre les distributions 'M' et 'F' tandis que la corrélation entre pathologies et ces mêmes marqueurs est mesurée par une ANOVA univariée selon les différentes pathologies du corpus.

À l'issue de ce calcul, nous ne conservons que les marqueurs ne corrélant pas avec une des caractéristiques précédentes : dans le cas contraire, les erreurs de transcription pourraient être dues non seulement à la somnolence, mais aussi aux caractéristiques en question. Nous obtenons alors des marqueurs qui sont *spécifiques* de la somnolence au regard des caractéristiques que l'on a mesurées.

Sensibilité à la somnolence : Test de Mann-Whitney Un deuxième prérequis pour identifier des biomarqueurs est leur *sensibilité* à la somnolence. Pour cela, nous sélectionnons pour chaque tâche uniquement les marqueurs étant significativement différents ($p < 0.05$) entre les deux distributions 'Somnolent' et 'Non somnolent' avec un test statistique de Mann-Whitney.

14.3.3 Résultats

Les marqueurs sélectionnés selon la procédure décrite ci-dessus sont décrits dans le tableau 14.3.

Tâche	Erreur		Système		MW		AUC
	Type	Unité	Unité	ML	U	p	
TILE	Nb. Sub.	mots	char	ester (char)	547.5	0.044	0.63
	% Sub.				550	0.046	0.62
ESS	Nb. Ins.	mots	char	ester (word)	658	0.02	0.64
	% Ins.				646.5	0.02	0.64
	Nb. Ins.	char			673.5	0.03	0.62
	% Ins.				673	0.03	0.63
KSS	% Ins.	mots	char	-	330	8.4×10^{-4}	0.54
	Nb. Ins.		char	ester (word)	398	0.047	0.63
	Nb. Ins.		char	affiné (char)	354.5	0.02	0.53
	% Ins.				357	0.017	0.54

TABLEAU 14.3 – Biomarqueurs de la somnolence élaborés à partir des erreurs de transcription de système de reconnaissance automatique de la parole. ML : Modèle de langage ; MW : Test de Mann-Whitney ; AUC : Aire sous la courbe ROC.

Deux tendances générales se dégagent de ces marqueurs :

— Le nombre d'insertions de mots ou de caractères semble lié au ressenti de somnolence des patients (ESS, KSS) ;

— Le nombre de substitutions, lui, semble refléter la somnolence telle que mesurée par des paramètres PSG, indépendamment du ressenti des patients.

Aucun de ces marqueurs ne semble, à lui seul, avoir assez de pouvoir discriminant pour être un biomarqueur efficace permettant la détection de la somnolence (AUC < 0.65), mais ce groupe de marqueurs reste prometteur pour compléter et améliorer les approches précédentes.

Les deux prochaines sections présentent deux approches inspirées de celle développée dans cette section pour deux tâches de détection de la somnolence au long cours :

— la détection de la SDE sévère (telle que mesurée par l'ESS, section 14.4) ;

— la détection de la propension à l'endormissement (telle que mesurée par la latence moyenne d'endormissement au TILE, section 14.5)

Pour cela, nous ajoutons aux présentes erreurs de STA les marqueurs acoustiques introduits dans le chapitre 11 et les erreurs de lecture, annotées manuellement, comme définies dans le chapitre 13.

14.4 Détection de la somnolence diurne excessive

14.4.1 Tâche investiguée et corpus

Dans cette section, nous nous concentrons sur la détection de la somnolence diurne excessive, telle que mesurée par l'ESS.

Dans la section précédente, un seuil de 10, correspondant au seuil classique de la présence de somnolence diurne excessive, a été utilisé. Cependant, les sujets inclus dans la base TILE sont des patients hypersomniaques, présentant pour la majorité des plaintes de SDE (79/93 sujets sont au-dessus de ce seuil). Ce seuil conduit à une répartition trop déséquilibrée pour permettre à un classifieur de généraliser correctement la tâche.

Dans un précédent article sur les performances de conduite, un score à l'ESS supérieur à 15 a été relié à des performances de conduite significativement dégradées par rapport au groupe non somnolent (ESS $<$ 10) et au groupe somnolent (11 $\leq$ ESS $\leq$ 15) (Philip *et coll.*, 2008). Ce seuil, correspondant à une plainte *sévère* de SDE, conduit à une répartition binaire plus équilibrée que le seuil précédent de 10 (cf tableau 14.4). Nous choisissons donc ce seuil dans cette section, afin d'investiguer la pertinence des descripteurs conçus jusqu'alors pour la détection de la plainte de SDE sévère chez des patients hypersomniaques.

En utilisant la base TILE-93, la répartition des patients entre les deux classes est représentée dans le tableau 14.4.

Sexe	ESS $>$ 15	ESS $\leq$ 15	TOTAL
F	26	32	58
M	13	22	35
TOTAL	39	54	93

Tableau 14.4 – Répartition des locuteurs dans les classes de somnolence en fonction de leur sexe.

14.4.2 Système proposé

Le système de classification proposé, inspiré des méthodologies de validation clinique d'outils psychométriques, est divisé en deux parties :

— d'une part, un processus de sélection des biomarqueurs, reprenant les conditions de spécificité et sensibilité à la somnolence telles que décrites dans la section précédente ;

— d'autre part, un classifieur composé d'une analyse en composantes principales (ACP) et d'une régression logistique.

La principale contrainte qui guide ces choix méthodologiques est la nécessité de faire un système composé uniquement d'outils statistiques et computationnels habituellement utilisés en recherche clinique, afin de faciliter la translation de ces outils dans une éventuelle pratique clinique.

Un schéma récapitulatif du système et des marqueurs sélectionnés à chaque étape du processus de sélection des marqueurs est proposé dans la figure 14.3.

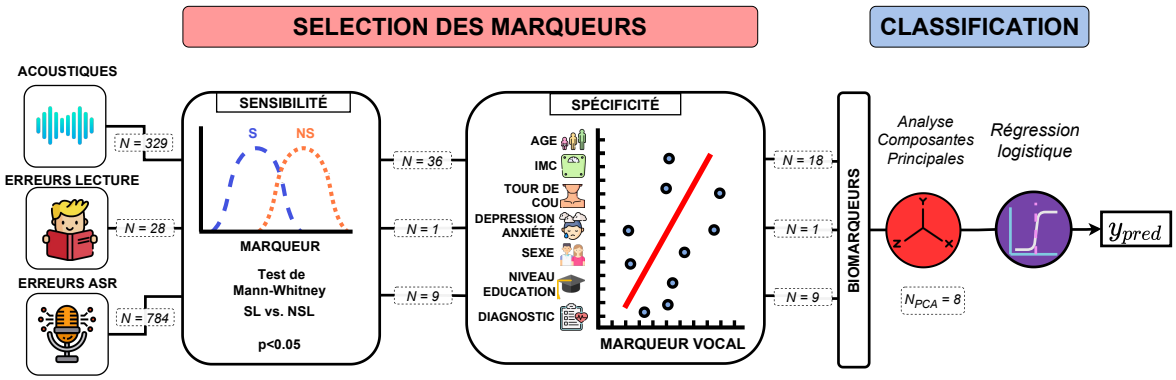

FIGURE 14.3 – Schéma du système proposé pour la détection de la plainte de SDE excessive. Les quantités en pointillées représentent le nombre de marqueurs à chaque étape du système.

Marqueurs Les marqueurs utilisés en entrée de ce système sont de trois natures :

— des marqueurs acoustiques (n=44), décrits dans le chapitre 11 ;
— les erreurs de lecture (n=4), décrites dans le chapitre 13 ;
— et les erreurs de STA (n=80), décrits dans la section précédente. Pour ce premier système, nous nous sommes restreints aux systèmes de STA ayant un modèle de langage.

Fusion précoce Dans le système proposé, nous faisons une fusion précoce des tous les marqueurs, mesurés sur les cinq siestes auxquelles nous ajoutons la moyenne et l'écart-type par locuteur sur les siestes. Cela conduit à un vecteur de 917 biomarqueurs potentiels de la somnolence.

Sensibilité et spécificité à la somnolence De même que dans la section précédente, nous sélectionnons uniquement les marqueurs qui sont à la fois *sensibles* et *spécifiques* à la somnolence telle qu'elle est définie ici, en ne conservant que les marqueurs ayant des distributions significativement différentes sur les deux classes, et en éliminant ceux corrélant avec un des cofacteurs mesurés.

Classification Enfin, le module de classification est composé d'une ACP, conservant 80% de la variance originale, suivie d'une régression logistique. Afin de prendre en compte le déséquilibre des classes ; cette dernière pondère les exemples d'entraînement par l'inverse du nombre d'échantillons dans chaque classe de somnolence.

Validation croisée Pour l'obtention des résultats de classification, nous avons fait une validation croisée des performances du système (LOSOCV, cf. chapitre 16). Seule la partie classification du système est validée ainsi : la sélection des marqueurs est effectuée sur l'entièreté du corpus. Même si cette pratique peut biaiser les performances obtenues, par le fait que le label est utilisé dans l'étape permettant de vérifier la sensibilité des marqueurs à la somnolence, cette procédure, fortement inspirée de la validation clinique des biomarqueurs, a pour but d'accepter ou de rejeter l'utilisation des erreurs des systèmes de transcription comme biomarqueurs de la SDE sévère. La classification proposée se place dans un contexte où l'on aurait une connaissance a priori des biomarqueurs de la somnolence (artificiellement déterminés ici lors de la sélection des descripteurs), à partir desquels nous souhaiterions entraîner un classifieur. Le résultat principal porte plus sur la nature des marqueurs sélectionnés que sur

les performances de classification en elles-mêmes. Pour un schéma de classification rigoureux (avec double validation croisée), nous invitons le lecteur à se reporter au chapitre 16.

14.4.3 Résultats

Performances de classification Les performances obtenues (UAR, moyenne pondérée des F1-scores et Aire sous la courbe ROC) pour différentes combinaisons de marqueurs sont présentées dans le tableau 14.5. Les courbes ROC correspondantes, mesurant le pouvoir discriminant des classifieurs, sont présentées dans la figure 14.4 (gauche).

	Descripteurs	UAR (%)	F1 (%)	AUC (%)
(a)	STA	63.5	59.5	65.8
(b)	Acoustiques	64.2	61.3	69.2
(c)	E. lecture	61.2	49.2	35.4
(d)	STA + Acoustiques	69.2	65.9	73.3
(e)	STA + E. lecture	64.5	60.2	66.2
(f)	Acoustiques + E. lecture	66.1	62.8	70.8
(g)	STA + Acoustiques + E. lecture	**74.2**	**70.7**	**78.6**

TABLEAU 14.5 – Performances de classification du système proposé pour la détection de la SDE sévère. *UAR* : *Unweighted Average Recall* ; *F1-score* : moyenne pondérée des F1-score, *AUC* : Aire sous la courbe ROC ; E. lecture : Erreurs de lecture, STA : Erreurs des systèmes de transcription automatique.

Lorsqu'ils sont pris séparément, les marqueurs acoustiques et les erreurs des systèmes de transcription conduisent à des performances identiques [systèmes (a) et (b), ≈ 64% d'UAR]. Cependant, leur combinaison surpasse les performances de toutes les combinaisons de deux marqueurs [69.2% pour le système (d)]. Au contraire, les erreurs de lecture conduisent à des performances faibles quand elles sont prises seules [61.2% d'UAR pour le système (c)] ou combinées à n'importe quel autre ensemble de marqueurs [64.5% d'UAR pour le système (e), 66.1% pour le système (f)]. Celles-ci semblent cependant indispensables à l'identification correcte de la SDE sévère : les meilleures performances sont obtenues en combinant les trois groupes de marqueurs et atteignent 74.2% d'UAR, 70.7% de F1-score et 78.6% d'AUC. La matrice de confusion correspondante est proposée dans la figure 14.4 (centre).

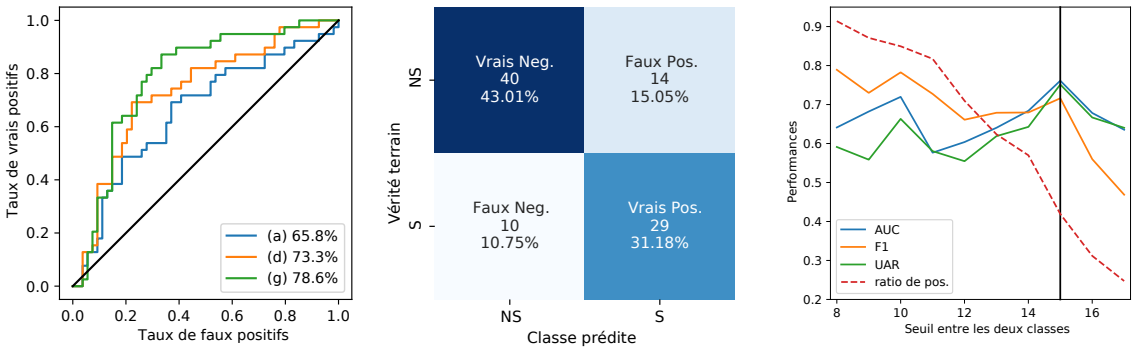

FIGURE 14.4 – Gauche : Courbes ROC des systèmes (a), (d), and (g) et leurs aires sous la courbe respective. Centre : Matrice de confusion du système (g). Droite : Performances du système (g) en fonction du seuil délimitant les deux classes Somnolent et Non Somnolent.

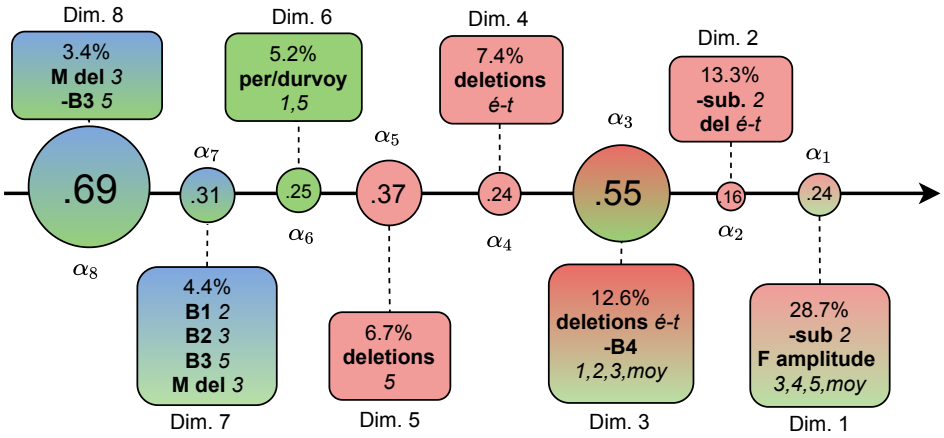

FIGURE 14.5 – Composantes de l'ACP et leur poids associés dans la régression logistique. De haut en bas : moyenne du ratio de variance expliquée dans l'ACP, descripteur (gras), modalité de mesure (en italique). Fond vert : descripteurs acoustiques ; fond bleu : erreurs de lecture ; fond rouge : erreurs des STA. - : poids négatif dans l'ACP ; *F* : formant ; *sub* : substitutions des STA ; *del* : délétions dans les STA ; *M del* : délétions annotées manuellement ; *Dim.* : Dimension.

Nécessité de l'ACP À première vue, le faible nombre de marqueurs passant la phase de sélection des marqueurs (28) devrait permettre d'entraîner le classifieur sans avoir besoin de technique de réduction de dimension telle que l'ACP. Pour tester cette hypothèse, nous avons réentraîné et ré-évalué le système (g) sans l'ACP. Cela conduit à des performances notablement plus faibles que le même système avec l'ACP (UAR : 68.2%, F1 : 63.3%, AUC : 69.7%). En effet, non seulement l'ACP est une technique de réduction de dimension, mais elle permet aussi d'orthogonaliser les marqueurs, optimisant l'apprentissage de la régression logistique.

Sensibilité à la limite entre les classes De la même façon que dans le chapitre 11, nous avons représenté dans la figure 14.4 (droite) les variations de performance (UAR) en fonction du seuil délimitant les deux classes durant la classification. Cette mesure reflète la spécificité des systèmes au seuil discriminant les deux classes lors de la classification binaire : les meilleures performances (UAR et AUC) sont observées pour une limite de 15, qui est celle sélectionnée précédemment. Un autre pic est observé pour une limite de 9, mais le déséquilibre entre les classes (moins 15% de négatifs) rend les conclusions périlleuses. En conclusion, notre système semble spécifique de cette valeur seuil sur l'ESS, qui correspond à une dimension de plainte sévère de SDE.

14.4.4 Analyse des marqueurs

Contribution des différents marqueurs

Au cours de la validation croisée, les paramètres de l'ACP et les poids de la régression logistique sont moyennés. La figure 14.5 représente les huit dimensions de l'ACP et leur poids correspondant dans la régression logistique.

Erreurs de lecture La dimension comptant le plus dans la régression logistique (Dim. 8, $\alpha_8 = 0.69$) est partiellement dirigée par les délétions annotées manuellement sur le troisième texte. Dans la septième et la huitième composante de l'ACP, ces erreurs de lecture ne sont

groupées qu'avec des marqueurs acoustiques : même si les erreurs de STA comptant dans la décision du classifieur comprennent des délétions, ces dernières ne semblent pas remplacer les annotations manuelles, mais plutôt être une nouvelle mesure de l'expression de la somnolence dans la voix, complémentaire des précédentes approches.

Marqueurs acoustiques Les marqueurs acoustiques sélectionnés par le classifieur sont principalement liés à la bande passante des formants :

— le troisième formant (B3) durant la cinquième sieste (Dim. 8) ;
— le quatrième formant (B4) durant la première, la deuxième et la troisième sieste et sa valeur moyenne au cours des siestes (Dim. 3, $\alpha_3 = 0.55$) ;
— le premier, deuxième et troisième formant resp. durant la deuxième, troisième et cinquième sieste (Dim. 7, $\alpha_7 = 0.31$) ;

L'amplitude des formants durant la troisième, la quatrième et la cinquième sieste et leur valeur moyenne au cours des siestes sont aussi pertinentes ($\alpha_1 = 0.24$).

Enfin, la durée et le pourcentage de durée des voyelles extraites des enregistrements audio durant les premières et cinquièmes siestes interviennent dans la décision du classifieur ($\alpha_6 = 0.25$).

Erreurs des systèmes de transcription automatique Concernant les erreurs des STA, les marqueurs les plus pertinents sont :

— l'écart-type des délétions au cours des siestes (Dim. 2, 3, 4) ;
— le nombre de délétions durant le cinquième sieste (Dim. 5) ;
— le nombre de substitutions durant la deuxième sieste (Dim. 1, 2).

Les erreurs qui ont été sélectionnées ont été produites par un STA avec un modèle de langage entraîné sur des caractères ou des mots, qui sont parmi les systèmes ayant les plus faibles taux d'erreurs (Boyer et Rouas, 2019).

14.5 Détection d'une propension à l'endormissement diurne pathologique

14.5.1 Tâche investiguée et corpus

Dans cette section nous nous intéressons à la détection d'une latence d'endormissement moyenne au TILE pathologique (S : TILE $\leq$ 8 min., NS : TILE $>$ 8 min.). En utilisant la base TILE-93 (cf. chapitre 7 et annexe E), la répartition des patients entre les deux classes est représentée dans le tableau 14.6.

Sexe	S	NS	TOTAL
F	10	48	58
H	11	24	35
TOTAL	**21**	**72**	**93**

Tableau 14.6 – Répartitions des locuteurs dans les classes de somnolence en fonction de leur sexe. S : Somnolent (TILE $\leq$ 8 min.), NS : Non-Somnolent (TILE $>$ 8 min.)

14.5.2 Système proposé

Le système proposé pour cette tâche est identique à celui proposé dans la section précédente, à la différence près que les seules erreurs des STA considérées dans cette section sont les insertions et les substitutions, en accord avec les résultats préliminaires de la section 14.3. Une représentation succincte du système est proposée dans la figure 14.6.

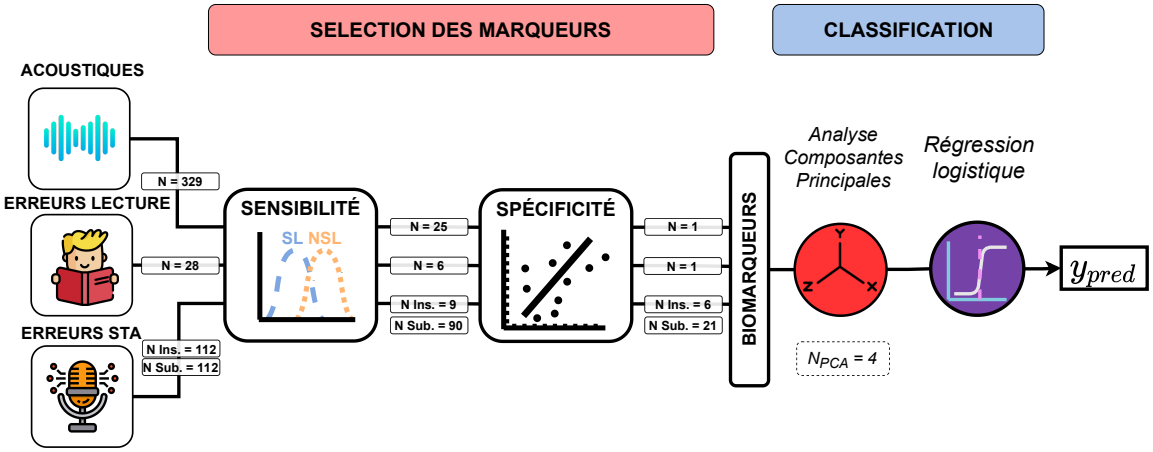

FIGURE 14.6 – Schéma du système proposé pour la détection de la propension à l'endormissement diurne. Les quantités encadrées représentent le nombre de marqueurs à chaque étape du système.

14.5.3 Résultats

Validation croisée

De la même façon que dans la section précédente, la validation croisée n'est effectuée que sur la partie de classification du système. Pour l'élaboration d'un classifieur rigoureux (double validation croisée), nous redirigeons le lecteur vers le chapitre 16.

Performance de classification

	Descripteurs	UAR (%)	F1 (%)	AUC(%)
(a)	STA	**73.2**	**75.8**	**74.8**
(b)	E. lecture	57.7	73.4	22.1
(c)	Acoustiques	59.5	66.0	60.1
(d)	STA + E. lecture	71.8	74.0	74.5
(e)	STA + Acoustiques	73.2	75.9	74.4
(f)	E. lecture + Acoustiques	61.3	70.2	67.7
(g)	Tous	73.9	76.8	74.6

TABLEAU 14.7 – Performances de classification du système pour la détection de la propension à l'endormissement diurne. *UAR* : *Unweighted Average Recall* ; *F1-score* : moyenne pondérée des F1-score, *AUC* : Aire sous la courbe ROC ; E. lecture : Erreurs de lecture, STA : Erreurs des systèmes de transcription automatique.

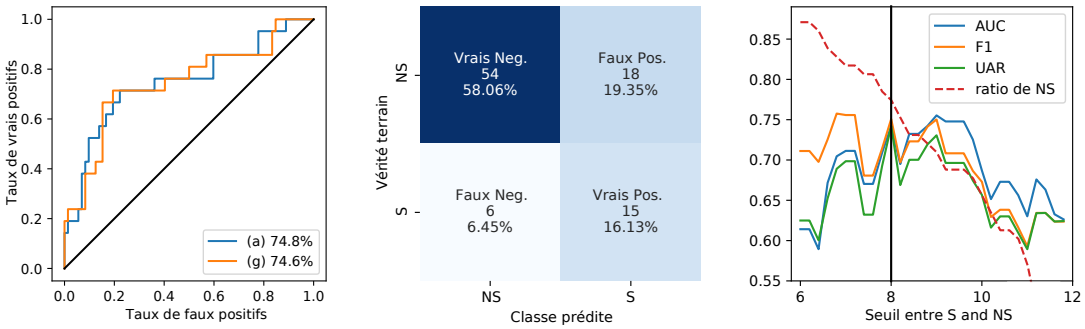

FIGURE 14.7 – Gauche : Courbe ROC des systèmes (a) et (g) et leurs aires sous la courbe correspondantes. Centre : Matrice de confusion du système (a). Droite : Performances du système (a) en fonction du seuil délimitant les deux classes Somnolent et Non-Somnolent.

Les performances du système proposé (UAR, moyenne pondérée des F1-score et Aire sous la courbe ROC) pour les différentes combinaisons de descripteurs vocaux sont présentées dans le tableau 14.7.

Les meilleures performances sont obtenues par le système (g), qui s'appuie sur les trois types de marqueurs : ce système atteint 73.9% d'UAR, 76.8% de F1-score et 74.6% d'aire sous la courbe ROC. Dans ce système, les erreurs sélectionnées sont le nombre d'additions lors de la lecture du quatrième texte, tandis que le marqueur acoustique sélectionné est la bande passante du premier formant sur la première sieste.

Cependant, il est à noter que les erreurs des STA seules [système (a)] permettent d'atteindre des performances de classification qui ne sont que légèrement plus faibles que celles du système (g) : 73.2% d'UAR, 75.8% de F1-score, et une aire sous la courbe ROC de 0.75. Le marqueur acoustique et l'erreur de lecture sélectionnés par le système semblent avoir peu d'importance dans la décision du classifieur.

En mettant en balance les bénéfices des erreurs de lecture annotées manuellement au regard du faible gain de performance qu'elles apportent (0.7% de gain absolu sur l'UAR, 1% sur le F1-score), nous choisissons de ne pas conserver le système (g). Puisque la combinaison des erreurs des STA et de marqueurs acoustiques [système (c)] produit des performances inférieures aux erreurs de STA seules [système (a)], nous choisissons de nous concentrer sur ce dernier.

Les courbes ROC des systèmes (a) et (g) ainsi que la matrice de confusion du système (a) sont représentées dans la figure 14.7. La proximité des deux courbes ROC confirme la similarité des systèmes (a) et (g) et consolide notre choix d'étudier le système (a).

Sensibilité à la limite entre les classes

De même que dans la section précédente, nous avons représenté dans la figure 14.7 (droite) les performances du système (a) en fonction du seuil délimitant les deux classes S et NS.

Comme attendu, les meilleures performances sont obtenues pour un seuil de 8 minutes. De plus, excepté un creux pour un seuil de 7.5 minutes, le système (a) permet des UAR de plus de 70% pour tous les seuils entre 7 minutes et 9.5 minutes, permettant aux médecins de sélectionner la sévérité de tendance à l'endormissement diurne qu'ils souhaitent détecter.

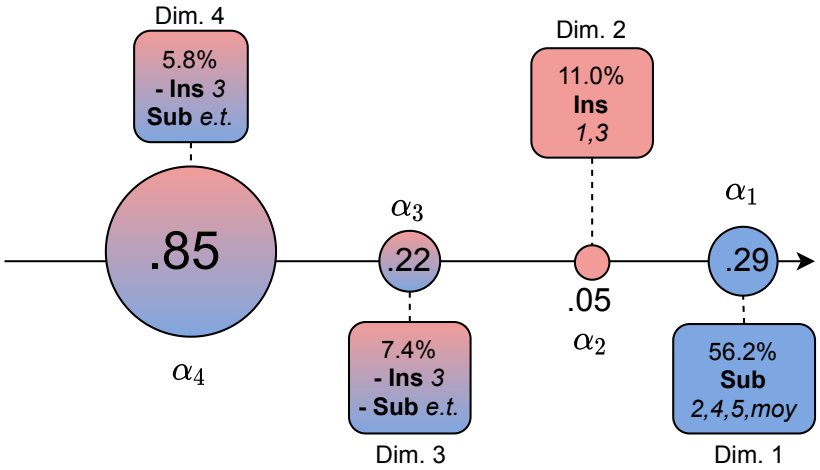

FIGURE 14.8 – Composantes de l'ACP et leur poids associé dans la régression logistique. De haut en bas : moyenne du ratio de variance expliquée dans l'ACP, descripteur (gras), modalité de mesure (en italique). Fond bleu : substitutions faites par les STA ; Fond rouge : insertions faites par les STA.

14.5.4 Analyse des marqueurs

Contribution des différents marqueurs

Les contributions des différentes erreurs des STA à l'ACP et des différentes composantes de l'ACP à la décision de la régression logistique sont représentées dans la figure 14.8.

La composante de l'ACP la plus importante dans la décision du classifieur est la dimension n°4 ($\alpha_4 = 0.85$), qui est dirigée par la différence entre le nombre d'insertions lors de la troisième session et l'écart-type, au cours des siestes, du nombre de substitutions. La somme des mêmes marqueurs anti-dirige la troisième composante de l'ACP, qui a un poids $\alpha_3 = 0.22$ dans la décision du classifieur.

Dans ces deux dimensions, les erreurs d'insertion durant la lecture du troisième texte sont faites par un système de transcription automatique basé caractères avec un modèle de langage basé mots, tandis que les substitutions sont faites par un système de transcription basé sur des portions de mots sans modèle de langage.

La deuxième dimension la plus importante ($\alpha_1 = 0.29$) est dirigée par le nombre et le ratio de substitutions faites par le système sur les enregistrements de la deuxième, quatrième et cinquième session, et leur valeur moyenne au cours des sessions. Ces erreurs proviennent de sept systèmes de transcription automatique basés sur différentes unités acoustiques, avec et sans modèles de langage.

Enfin, la dimension la moins importante dans la décision ($\alpha_2 = -0.05$) est dirigée par le nombre d'insertions sur les premiers et troisièmes enregistrements. Contrairement aux insertions qui avaient été sélectionnées sur la troisième session, les insertions de la première session sont faites par un système basé sur des BPE, sans modèle de langage.

14.6 Conclusion et perspective

Les résultats présentés dans ce chapitre ouvrent des perspectives de recherche sur les mécanismes explicitant l'impact de la somnolence sur ces marqueurs. Une approche utilisant des techniques d'intelligence artificielle à visée explicative (*explainable AI*) pourrait permettre

de lier les mécanismes internes des STA utilisés aux variations de la voix dues à la somnolence et ainsi contribuer à l'explication du mécanisme sous-jacent liant somnolence et production vocale.

À l'inverse, l'étude de l'influence des caractéristiques du locuteur sur les erreurs des STA pourrait permettre l'amélioration de ceux-ci.

Ces questions de recherche sont laissées ouvertes, le chapitre suivant introduisant une autre dimension importante de la parole : la respiration et le rythme de lecture, et plus particulièrement les pauses lors de la lecture à voix haute.

Chapitre 15

Pauses de lecture

Sommaire

Les travaux présentés dans ce chapitre ont bénéficié du travail de Brice Arnaud (BA), accueilli dans le cadre de son stage de L3 à l'Université de Bordeaux (2021), et d'Agathe Basse (AB), de Benoît Caudron (BC) et de Marie Huillet (MH), accueillis dans le cadre de leur stage de recherche pour leur M1 d'orthophonie au CFUOB (2021).

15.1 Contexte et motivations

Nous avons vu dans les deux chapitres précédents que les erreurs de lecture (et leur équivalent pour les systèmes de transcription automatique) sont des marqueurs pertinents pour l'estimation de la somnolence à partir de marqueurs vocaux, et plus particulièrement à travers l'impact de la somnolence sur les capacités cognitives du sujet.

Ce chapitre propose un nouveau jeu de descripteurs du niveau de capacité de lecture en lien avec la somnolence évaluant le phrasé de la voix des patients, et plus précisément, les pauses de lecture entre groupes de mots. En effet, alors que certains patients ont un flux de lecture correct, en s'arrêtant là où cela est naturel, d'autres augmentent la longueur des pauses afin de planifier la suite de la lecture, hésitent, ou se corrigent après avoir réalisé qu'ils ont commis une erreur. Cela conduit à des changements de flux de lecture, qui reflètent leurs difficultés dans la planification cognitive.

Des travaux récents sur la maladie d'Alzheimer (Balagopalan *et coll.*, 2021; Bose *et coll.*, 2021; Clarke *et coll.*, 2021; Gonzalez-Atienza *et coll.*, 2021; Haulcy et Glass, 2021), la sclérose en plaques (Noffs *et coll.*, 2020), l'anxiété (Albuquerque *et coll.*, 2021; Kim *et coll.*, 2020), et même la dépression (Demiroglu *et coll.*, 2020; Zhang *et coll.*, 2020) se sont déjà intéressés à des marqueurs dérivés des pauses lors de la lecture ou de la parole spontanée (cf Chapitre 2).

Cependant, à notre connaissance, aucun travail publié n'a étudié l'emplacement de ces pauses dans le discours ou le texte original. L'originalité des travaux présentés dans ce chapitre repose sur la conception de biomarqueurs vocaux de la somnolence liés non seulement à la durée des pauses de lecture, mais aussi à leur emplacement dans le texte original.

Dans la section 15.2, nous présentons le système mis au point pour l'extraction automatique des durées et emplacements des pauses de lecture. Dans la section 15.3, nous décrivons et discutons la fiabilité d'une annotation théorique de la naturalité des pauses de lecture dans les textes de référence. Enfin, dans la section 15.4, nous décrivons les marqueurs extraits à partir du système automatique et des annotations théoriques, et analysons les profils de lecture dérivés de ces marqueurs pour les relier aux caractéristiques des locuteurs.

15.2 Extraction automatique des durées et emplacements des pauses de lecture

Dans un premier temps, nous extrayons de manière automatique les durées et les emplacements des pauses de lecture à partir des enregistrements audio des patients. Cette procédure est effectuée en trois étapes :

1. Conception d'un système de détection d'activité vocale – section 15.2.1 ;
2. Correction de l'hypothèse de sortie d'un système de transcription automatique avec le précédent système – section 15.2.2 ;
3. Alignement entre l'hypothèse corrigée du système de transcription automatique et le texte de référence – section 15.2.3 ;

15.2.1 Conception d'un détecteur d'activité vocale

Objectif

Cette section introduit l'algorithme mis au point pour corriger l'hypothèse de sortie d'un système de transcription automatique.

En effet, une première version de ce système reposait sur la sortie brute d'un système de transcription automatique, tel que décrit dans la section 15.2.2. Cependant, comme représentées dans la figure 15.1, les délimitations des zones de parole ou de silences sont imprécises, commençant trop tôt ou finissant trop tard. De plus, des petites zones de silence sont incorrectement identifiées au milieu de zones de parole, ce qui interfère grandement avec l'estimation du nombre de pauses.

L'objectif est donc d'obtenir une transcription automatique alignée de manière très précise avec l'audio afin d'avoir une estimation précise de la durée des pauses, tout en conservant le contenu lexical afin de pouvoir les situer dans le texte de référence.

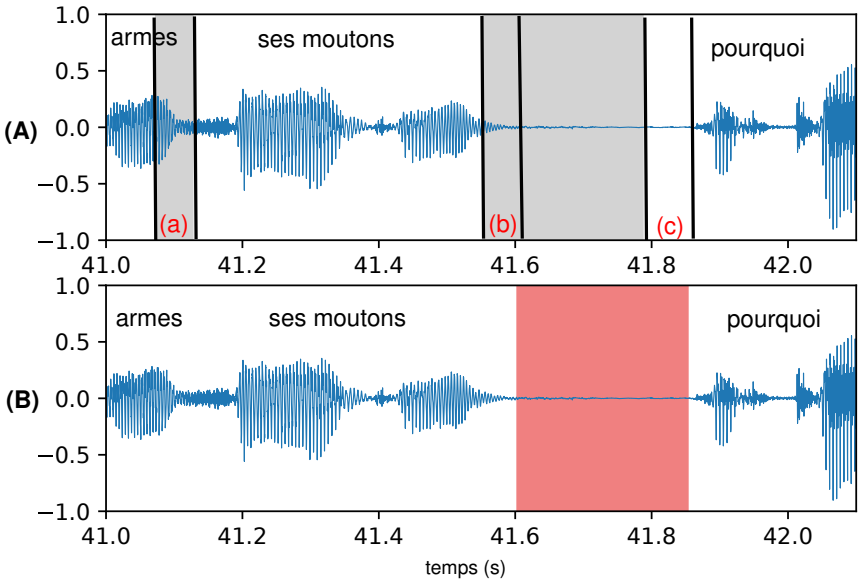

FIGURE 15.1 – **(A)** Hypothèse de transcription avant l'alignement par un système de détection d'activité vocale (DAV). Gris : segments de silences tels que détectés par le système de transcription automatique de la parole. (a) Segment de silence détecté incorrectement au milieu d'un segment de parole. (b) Début prématuré du segment de silence. (c) Début prématuré du segment de parole.
(B) Hypothèse de transcription après alignement avec notre système de DAV. Rouge : annotation des silences corrigée. Tous les segments de parole qui chevauchent les segments de parole détectés par le système de DAV sont concaténés dans le même segment final.

Détecteur d'activité vocale

Un détecteur d'activité vocale (DAV) est un algorithme dont le but est d'annoter un échantillon audio avec deux labels : « Parole » et « Silence ».

Cette problématique n'est pas récente, et de nombreux détecteurs d'activité vocale existent déjà et sont disponibles. Cependant, ceux-ci requièrent généralement de fixer des paramètres, et ce pour l'intégralité de la base de données (usuellement un seuil d'énergie au-dessus duquel le signal est considéré comme de la parole) [par ex. (Chen *et coll.*, 2020)] , et sont donc

peu appropriés pour traiter une base de données contenant des environnements acoustiques très variables, comme c'est le cas de la base TILE. D'autres systèmes, comme celui proposé par Google (2021), sont conçus pour faire de la détection d'activité vocale en temps réel et ne sont pas assez précis pour une estimation précise des pauses de lecture (résolution de l'ordre de 250 ms). En conséquence, nous avons conçu et évalué notre propre système de détection d'activité vocale.

Prétraitement Tous les fichiers sont prétraités avec la même procédure. Tout d'abord, nous rééchantillonnons tous les fichiers à la fréquence d'échantillonnage de 16 kHz et une résolution de 16 bits. Ensuite, nous appliquons l'algorithme de débruitage `N-HANS` (Liu *et coll.*, 2021), conçu pour débruiter les fichiers audio enregistrés dans des conditions réelles. Les enregistrements faits en conditions cliniques sont souvent bruités par des bruits de ventilation ou des bruits de pas. Pour les prendre en compte lors du traitement des données, nous avons enregistré le corpus TILE de façon à ce que chaque enregistrement contienne 500 ms de bruit environnemental. C'est cette portion d'enregistrement qui est utilisée comme référence de bruit avec `N-HANS`. Enfin, nous normalisons le volume sonore (*loudness*) de tous les fichiers à -3dB.

Détermination automatique du seuil d'énergie

Calcul de l'énergie La première étape de la détermination des zones de parole est le calcul de l'énergie du signal E (en dB), sur une fenêtre de taille W_{nrj}. L'énergie d'un signal s sur une fenêtre w de taille n est définie par :

$$E[w] = \frac{1}{n}\sqrt{\sum_{i=0}^{n} s[i]^2}$$
$$E[w]_{dB} = 10\log_{10}\left(E[w]\right)$$

Ensuite, nous déterminons automatiquement un seuil d'énergie, au-dessus duquel, pour chaque fenêtre d'analyse, nous décidons si la fenêtre contient de la voix ou non. Nous établissons un algorithme adaptatif, qui permet de choisir le meilleur seuil pour chaque échantillon, détaillé ci-après.

Moyenne des plus grandes différences d'énergie Parmi toutes les méthodes proposées par BA pour élaborer un seuil adaptatif pour différencier parole et bruit de fond sur les différents environnements acoustiques de la base TILE, nous avons retenu une méthode basée sur la moyenne des plus grandes différences d'énergie. Pour cela, nous procédons de la manière suivante :

1. Nous calculons $\Delta_i = |E[i+\delta] - E[i]|$, la différence entre deux fenêtres d'énergie séparées de δ ;
2. Ensuite, nous trions les Δ_i et gardons les N_{peaks} plus hautes valeurs ;
3. Enfin, pour chacun des précédents Δ_j^{kept}, nous calculons m_j, l'énergie médiane de l'intervalle sur lequel Δ_j^{kept} a été calculé.

Le seuil choisi pour le fichier audio considéré θ est ensuite défini comme la moyenne des m_j, tandis que δ est un hyperparamètre du système. Cette méthode est illustrée dans la partie gauche de la figure 15.2.

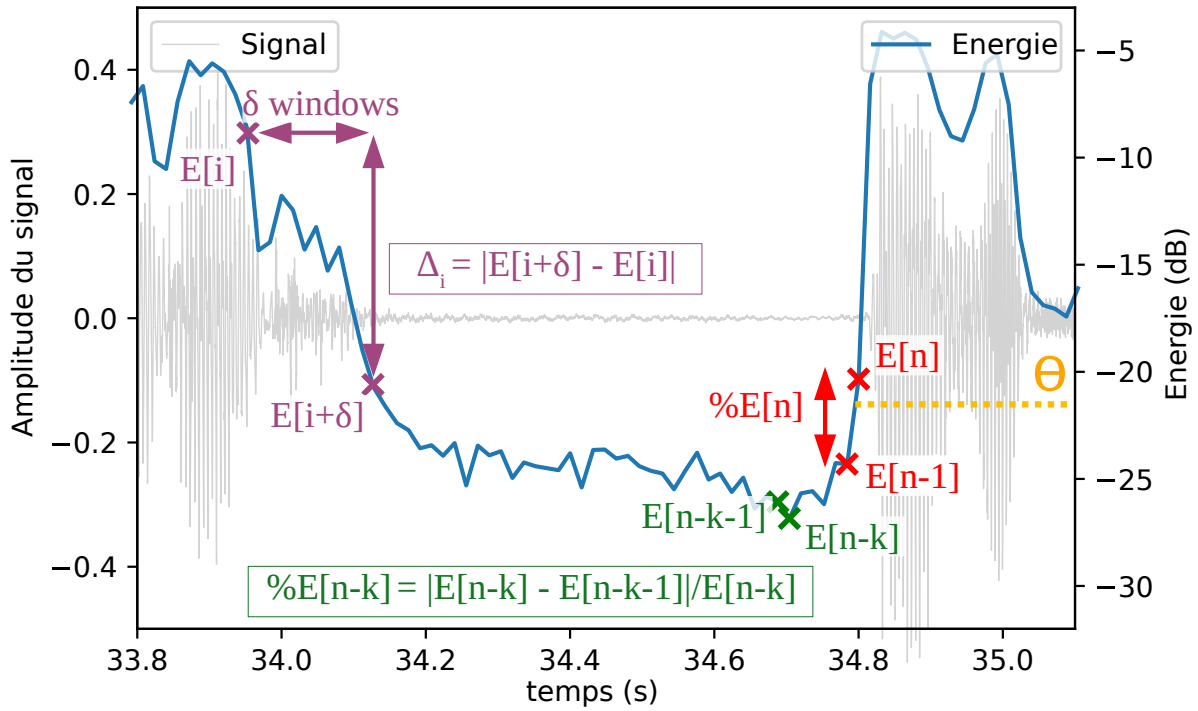

FIGURE 15.2 – Gauche : exemple de calcul de Δ_i afin de calculer le meilleur seuil θ. Droite : en notant n la première fenêtre pour laquelle $E[n] > \theta$, le début du segment de parole est défini par le plus petit k tel que $\%E[n-k] < \alpha$. Exemple tracé pour $k = 6$, $\alpha = 0.1$ et $\theta = -21$dB.

Début et fin des segments de parole Le principal inconvénient de l'utilisation d'un système de DAV basé sur un seuil d'énergie est que les parties voisées commencent souvent avant que l'énergie dépasse θ (cf. partie), et inversement. Afin de prendre cela en compte, nous avons ajouté de l'inertie au changement entre les classes détectées (Parole ou Silence). Ainsi, nous définissons le début d'un segment de parole comme ceci :

1. Nous calculons la première fenêtre n pour laquelle $E[n] > \theta$;
2. Pour $k \in \{0, 1, 2, ...\}$, nous calculons $\%E[n-k]$, le gain relatif d'énergie entre la fenêtre $n-k-1$ et la fenêtre $n-k$;
3. La première fenêtre $n-k$ pour laquelle $\%E[n-k] < \alpha$ est considérée comme le début de la zone de parole.

Cette procédure est illustrée dans la partie droite de la figure 15.2. Une procédure similaire est appliquée pour détecter la fin des segments de parole.

Enfin, notre but est de concevoir des marqueurs reflétant le comportement de lecture des patients. Les courts bruits de fond ou les bruits de bouche peuvent déclencher notre DAV. Pour éviter ce comportement, nous filtrons les segments de voix qui sont plus courts que le seuil η_{voix}. De même, il est naturel de prendre de courtes pauses lorsqu'on lit un texte à voix haute, ce qui se traduit par de très courts segments de silences lors de l'analyse avec le DAV. De la même manière que pour les segments de voix, nous avons filtré les silences qui sont plus courts que le seuil η_{sil}.

Recherche des paramètres

Afin de trouver les meilleurs paramètres pour notre système de DAV, BA a annoté manuellement les segments de silence et de parole de fichiers audio extraits du corpus TILE. Ensuite, nous calculons les performances des différents systèmes de DAV et identifions le meilleur jeu de paramètres à appliquer à l'entièreté de la base de données.

Métrique La métrique utilisée dans cette section est la valeur moyenne de $\cap_i$, la durée d'agrément entre la vérité terrain et l'annotation estimée sur le fichier i, pondéré par la longueur de chaque fichier audio :

$$\cap_i = l^i_{\text{agréément}} / l_{\text{i}}$$

Annotation de 25 fichiers Afin d'estimer les meilleurs paramètres de notre système de DAV, BA a annoté manuellement les segments de parole et de silence de 25 échantillons tirés pseudo-aléatoirement dans le corpus TILE (durée moyenne : 77.7 ± 11.3s, nombre moyen de pauses : 30.1 ± 7.3 pauses, temps total moyen de pauses : 17.6 ± 7.7s). Les inspirations ont été annotées séparément.

Validation croisée L'évaluation de toutes les valeurs possibles pour les paramètres est faite avec une validation croisée de 5 fois 5-*fold* : les échantillons sont aléatoirement distribués en cinq groupes, qui servent tour à tour de base d'évaluation, tandis que les 4/5 restants servent de base d'entraînement. La procédure est répétée cinq fois, puis les performances sont recalculées une dernière fois avec la médiane des paramètres donnant les meilleures performances sur chaque itération.

Résultat Cette procédure a conduit à un score de concordance pondérée final de $\bar{\cap}_{\text{DAV}} = 93,2\%$ pour l'hypothèse du STA alignée sur le DAV avec les paramètres décrits dans le tableau 15.1, ce qui représente une amélioration significative de près de 1% par rapport au même score calculé sur la sortie brute du STA ($\bar{\cap}_{\text{ASR}} = 92,3\%$, test de Mann Whitney : $p = 0.03$, STA présenté dans la prochaine section).

De plus, afin prendre en compte le nombre de pauses identifiées par chaque système, nous calculons l'erreur absolue moyenne (MAE) entre le nombre de pauses annotées manuellement et celles détectées par le STA brut et le système aligné sur le DAV. La MAE_{STA} entre le nombre de pauses identifiées dans l'hypothèse brute du STA et la vérité terrain atteint 24.12 pauses, tandis que son homologue aligné sur le DAV identifie plus précisément le nombre de pauses ($\text{MAE}_{\text{aligné}}$ = 4.24 pauses).

Cela valide notre utilisation de ce système de DAV.

15.2.2 Correction de l'hypothèse de transcription

Système de transcription automatique

Les systèmes de transcription automatique bout-en-bout introduits dans le chapitre 14 ne sont pas contraints temporellement, et ne permettent donc pas l'alignement entre l'hypothèse de transcription et le texte de référence. Nous avons donc choisi un modèle chaîné pour extraire les silences entre groupes de mots, implémenté dans la boîte à outils `kaldi` (Povey *et coll.*, 2011).

Ce modèle est un TDNN-HMM entraîné avec la fonction d'objectif LF-MMI. Le réseau de neurones est basé sur un réseau à délai (*time-delay neural network* – TDNN) avec 7 couches de

Nom	Val. (WSR)
W_{nrj}	{1024, **512**}
N_Δ	{**100**, 200}
δ	{**5**,8,10}
α	{0.05, **0.1**, 0.15, 0.20}
$\eta_{sil.}(ms.)$	{**200**, 250}
$\eta_{voix}(ms.)$	{200, **250**}

Tableau 15.1 – Paramètres optimaux pour le système de DAV.

TDNN comportant 1024 unités chacune, le pas temporel étant fixé à 1 dans les trois premières couches, 0 dans la quatrième couche et 3 dans la dernière couche.

Les entrées du modèle acoustique sont des MFCC (*Mel Frequency Cepstral Coefficient*) de haute résolution (40 dimensions), concaténés avec des i-vecteurs de dimension 100 (Gupta *et coll.*, 2014).

Le modèle de langage est un 3-gram basé sur les mots entraînés en utilisant la méthode de comptage de SRILM (Stolcke, 2002).

Le lexique utilisé est le dictionnaire phonétique fourni par le LIUM. Le dictionnaire est limité aux 50,000 mots les plus fréquents existant à la fois dans les textes d'entraînement et dans le dictionnaire.

Le système entier de transcription atteint un taux d'erreur de mots de 13.7%. Pour plus d'information sur l'implémentation de ce système de transcription, nous redirigeons le lecteur vers la thèse de Florian Boyer dont est issu ce système de transcription automatique (Boyer, 2021).

Alignement entre la transcription et le DAV

Afin d'aligner la sortie du système de transcription avec le système de DAV, nous concaténons tous les mots de l'hypothèse de transcription pour lesquels les bornes temporelles chevauchent les segments de voix identifiés par le système de DAV.

Ensuite, nous réassignons les bornes temporelles de ces segments dans l'hypothèse de transcription avec celles identifiées par le système de DAV. Un exemple de réalignement est représenté dans la figure 15.1, au début de cette section.

15.2.3 Alignement entre l'hypothèse de transcription corrigée et le texte de référence

La dernière étape pour extraire les emplacements des pauses dans le texte de référence est l'alignement entre l'hypothèse de transcription corrigée et le texte de référence.

Pour cela, nous appliquons la procédure suivante :

1. Nous calculons la distance de Smith-Waterman (Smith et Waterman, 1981) entre chaque segment de parole de l'hypothèse de transcription et tous les segments du texte original de taille $N_{texte} = N_{hyp.} \pm 2$, afin de prendre en compte les possibles fusions ou dissociations de mots dans l'hypothèse de transcription. Cette distance a été choisie en particulier car elle permet de calculer une distance directement entre les segments de manière globale, indépendamment des petites différences d'écriture qui peuvent

arriver entre la transcription et le texte de référence (pluriels, conjugaison des verbes homonymique ...) ;

2. Nous alignons ensuite les segments de l'hypothèse de transcription avec la position minimisant la distance précédemment calculée. Nous nous limitons à un intervalle de plus ou moins 15 segments autour de la fin du précédemment segment aligné, afin de contraindre la position des petits mots comme les prépositions ou les déterminants, qui peuvent avoir plusieurs positions optimales.

Le tableau 15.2 montre un exemple d'alignement entre l'hypothèse de transcription corrigée et le texte de référence.

Transcription automatique
je crois-tu qu'il **fait** beaucoup **d'armes** ∅ **ses** mouton \fɛ\ **d.**aʁm\ \se\
Texte de référence aligné
Crois-tu qu'il **faille** beaucoup **d'herbe** **à** **ce** mouton ? \faj\ \d.ɛʁb\ \sə\

TABLEAU 15.2 – Exemple d'alignement entre l'hypothèse de transcription corrigée et le texte de référence. Gras : différences entre les deux versions.

15.3 Annotation des textes

15.3.1 Objectifs

Comme mentionné précédemment, les études portant sur les pauses de lecture n'utilisent que l'estimation de la longueur des pauses (généralement grâce à un détecteur d'activité vocale) et leur nombre. Notre objectif est d'étudier, sur de la lecture oralisée et dans le cadre de la somnolence pathologique, la pertinence de nouveaux marqueurs : l'emplacement des pauses de lecture. Les sujets s'arrêtent-ils plutôt à des endroits où la pause est « naturelle » (comme au niveau d'un signe de ponctuation) ou non (par exemple en milieu d'un mot), et dans quelle proportion ?

Dans la section précédente, nous avons élaboré un système permettant l'extraction automatique des durées et des emplacements des pauses de lecture dans le texte de référence. Nous désirons maintenant construire une grille de notation de référence de ces pauses, afin de quantifier leur naturalité. Pour cela, nous avons demandé à AB, BC et MH d'annoter en triple aveugle les pauses entre chaque paire de mots consécutifs de chaque texte de notre corpus avec un score indiquant si une pause de lecture à cet emplacement leur semble naturelle ou non.

Comme tout processus d'annotation, évaluer la naturalité des pauses de lecture d'un texte reflète la subjectivité et la sensibilité propre des annotateurs à ce qu'est une pause « naturelle ». Afin de faire émerger les points communs et différences d'appréhension de ce concept, nous leur avons donc demandé, après avoir finalisé leurs annotations, de rédiger un rapport sur leur façon d'annoter. L'originalité de l'approche proposée dans cette section est de proposer une comparaison entre, d'une part l'évaluation subjective de chaque annotateur concernant sa façon d'annoter, et d'autre part des statistiques sur les annotations résultantes. Cette comparaison permet d'objectiver – ou non – les ressentis des annotateurs et d'ainsi évaluer la fiabilité de leur production pour une utilisation en tant que vérité terrain dans un système automatisé.

Les annotations sont ainsi évaluées selon deux axes :
— l'analyse des méthodes d'annotation entre les trois annotateurs, présentée dans la section 15.3.3 ;
— l'analyse des lieux de désaccords, présentée dans la section 15.3.4.

L'utilisation ultérieure de ces annotations dans un système automatisé est cependant conditionnée à un bon agrément interannotateurs sur les textes proposés. La corrélation intraclasse pour chaque combinaison d'annotateurs et chaque texte est présentée dans la section 15.3.5, et permet de conclure quant à l'éventuelle utilisation de ces annotations comme vérité terrain pour la suite.

15.3.2 Consigne d'annotation

AB, BC et MH ont annoté les textes de référence de la base TILE avec la consigne suivante :

> « Annotez à quel point il est naturel de faire une pause dans la lecture au milieu de chaque paire de mots consécutifs, -10 étant très peu naturel, 10 étant très naturel, 0 étant neutre ».

Il a été convenu de noter de 2 en 2, les niveaux de notation disponibles étant les suivants : -10, -8, -6, -4, -2, 0, 2, 4, 6, 8, 10. Les annotations ont été faites en triple aveugle, afin de pouvoir prendre en compte la définition propre à chaque annotateur du caractère «naturel» de chaque pause. Les annotations proposées ainsi que leur moyenne et leur écart-type sont disponibles en ligne [1]. Un extrait en est proposé dans le tableau 15.3.

MH	J'	-10	ai	-10	beaucoup	-10	vécu	-8	chez	-10	les	-10	grandes	-10	personnes.	8
AB	J'	-10	ai	-10	beaucoup	-10	vécu	0	chez	-10	les	-10	grandes	-10	personnes.	10
BC	J'	-10	ai	-10	beaucoup	-10	vécu	-8	chez	-10	les	-10	grandes	-10	personnes.	10
Moy	J'	-10	ai	-10	beaucoup	-10	vécu	-5.33	chez	-10	les	-10	grandes	-10	personnes.	9.33
É-t	J'	0	ai	0	beaucoup	0	vécu	3.77	chez	0	les	0	grandes	0	personnes.	0.94

Tableau 15.3 – Exemple d'annotation sur le Texte n°0 extrait du Petit Prince (cf. Annexe D pour le texte intégral).

15.3.3 Stratégies d'annotation

Évaluation subjective des annotateurs

Puisque l'annotation des pauses est un processus comprenant une grande part d'évaluation subjective, nous avons demandé à AB, BC et MH de décrire, a posteriori, leur façon d'annoter les pauses. Un condensé de ce rapport est proposé dans la figure 15.3.

Une première remarque concerne un accord sur la sanction par une note de -10 de la séparation des groupes nominaux, insécables. Ensuite, d'un point de vue global, les notes positives semblent favoriser les arrêts aux signes de ponctuation, tandis que les notes négatives reflètent des erreurs qui gênent le rythme de lecture ou qui conduisent à des erreurs de sens. Par ailleurs, de nombreux niveaux sont utilisés non pas de manière discrète, mais de manière groupée, en utilisant différentes nuances pour le même type de pause suivant le contexte dans lequel elle se fait. Par exemple, BC nous renseigne qu'« en fonction des proportions des différents groupes [il a] marqué d'un -8 ou d'un -6 une potentielle pause entre le sujet et son verbe, comme entre le verbe et son adverbe ou complément : de telles pauses introduisent des ruptures inégalement inappropriées. »

1. https://zenodo.org/record/5813261

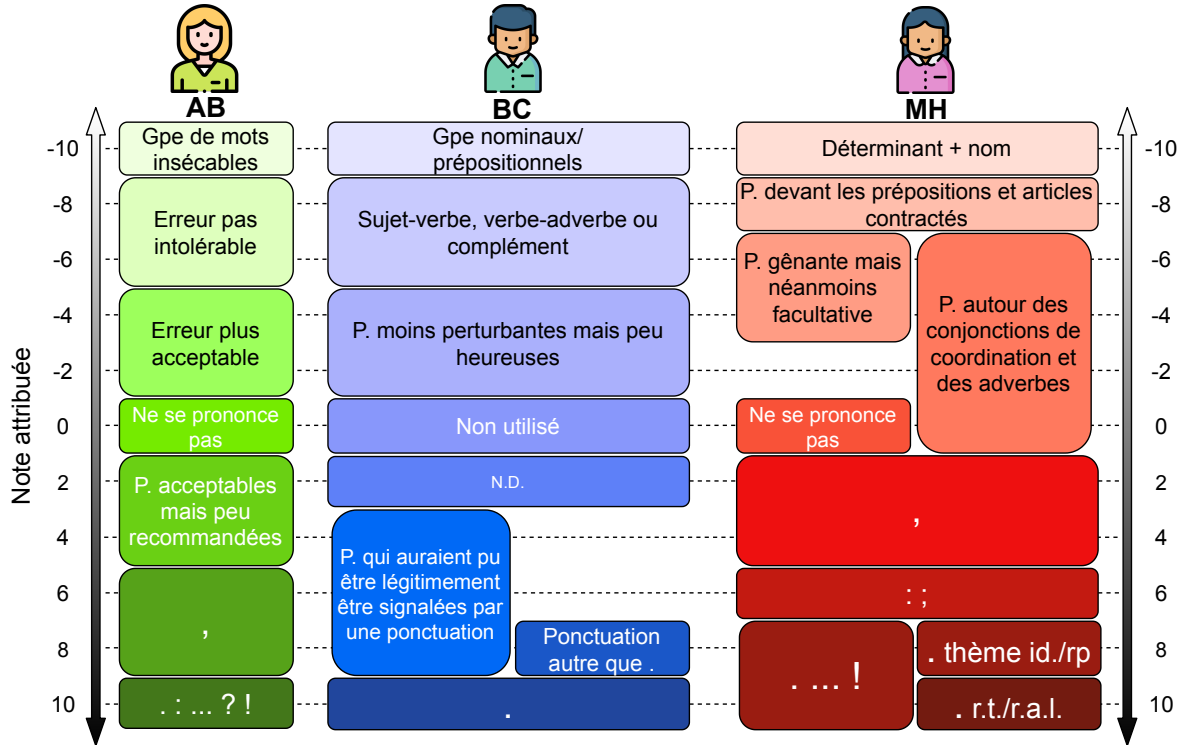

FIGURE 15.3 – Schéma illustrant les stratégies de chaque annotateur pour l'annotation de la naturalité des pauses de lecture dans les textes proposés. *Gpe* : Groupe ; *P* : Pauses ; *N.D.* : Non décrit ; *thème id.* : thème identique ; *rp* : reprise pronominale ; *r.t.* : rupture temporelle ; *r.a.l.* : retour à la ligne.

Concernant les pauses liées à la ponctuation, BC et AB sont globalement d'accord pour favoriser des pauses par les signes de ponctuation, quels qu'ils soient. MH se distingue d'une part par son traitement des points, qu'elle divise en deux classes suivant les groupes qu'ils séparent ; et d'autre part par son traitement des virgules auxquelles elle attribue des notes plus basses (+2 ou +4) que ses homologues. Elle explique cela par la façon dont elle parle couramment : « je ne pense pas marquer systématiquement les virgules par de réelles pauses, mais plutôt par des variations dans l'intonation. »

Enfin, il est à noter que BC précise n'utiliser que très rarement la note de 0 : « à mon sens, une pause est acceptable (positive) ou ne l'est pas (négative), quel que soit par ailleurs le degré d'acceptabilité. »

Approche statistique

Afin d'objectiver ces tendances d'annotations, nous avons représenté dans la figure 15.4 les histogrammes des annotateurs, tous textes confondus.

Ces histogrammes permettent de mettre en lumière un usage principal de la note -10 : sur le total des 1436 pauses annotées, plus de la moitié sont aberrantes (score de -10) et plus des trois quarts sont jugées comme n'étant pas naturelles (score négatif) par les annotateurs. Ceci est cohérent avec le comportement de lecture qu'aurait un sujet sain, qui ne s'arrêterait que quelques dizaines de fois (au niveau de pauses notées positivement) dans un texte contenant entre 200 et 250 mots.

Ces représentations permettent également de mettre en lumière trois distributions différentes des annotations. Chez BC, nous constatons une forte polarisation autour des notes -10

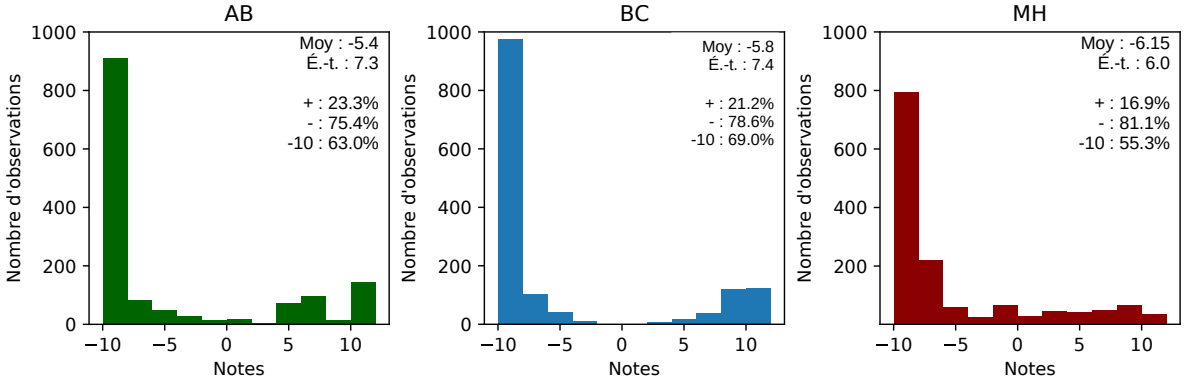

FIGURE 15.4 – Histogrammes des notes proposées par les annotateurs, tous textes confondus. *Moy* : Moyenne, *É.-t.* : Écart-type, + : ratio de scores positifs, - : ratio de scores négatifs.

et +8/+10, conséquence de sa conception des pauses de lecture : « [...] j'ai attribué la note de 10 à chaque pause faisant suite aux différents points possibles, ne serait-ce que pour sanctionner la règle – très élémentaire – selon laquelle une phrase commence par une majuscule et finit par un point. »

Nous retrouvons également une distribution partagée entre les deux pôles -10/+10 chez AB, mais qui est nuancée par une utilisation décroissante des niveaux intermédiaires au fur et à mesure que l'on s'approche de la notation neutre de 0. Enfin, si l'on retrouve un pôle -10/-8 chez MH, les notes positives sont utilisées de manière quasiment uniforme. Cela est une volonté de sa part : « Afin d'aboutir à une notation utilisant autant que possible tous les degrés présents dans l'échelle de notation et prenant en compte tous les découpages de phrase possibles, j'ai lu plusieurs fois des passages du texte à haute voix, en faisant varier mon rythme de lecture et mon intonation. »

Les différences observées entre les trois distributions sont statistiquement significatives (test de Kruskal-Wallis, $H = 14.7, df = 2, p = 6.5 \times 10^{-4}$). Un test post-hoc de Mann-Whitney sur chaque paire d'annotateurs permet de révéler une différence significative entre les annotations d'AB et de BC (MW, $p = 0.04$) et entre celles de MH et de BC (MW, $p = 3.9 \times 10^{-4}$), mais pas entre celles de MH et AB (MW, $p = 0.11$)

15.3.4 Lieux de désaccords

Évaluation subjective des annotateurs

Les six textes ont des caractéristiques différentes – longueur de phrases, complexité de la ponctuation, dialogues – qui peuvent influencer l'annotation des pauses de lecture. Nous avons ainsi identifié dans le rapport d'annotation quatre catégories de désaccords dont nous présentons des exemples prototypiques dans le tableau tableau 15.4.

(a) Dialogues : Dans son rapport, MH précise qu'elle a « rencontré plus de difficultés pour l'annotation des dialogues ». En effet, en raison de la double oralité de la lecture de ces passages à voix haute et de la mise en forme particulière, il peut être difficile d'estimer si une pause est naturelle ou non à la lecture.

(b) Virgules : Au-delà du fait que MH annote les virgules différemment de ses confrères, elles peuvent, suivant le contexte ou les caractéristiques de lecture de l'annotateur, être source

de désaccord entre les annotateurs.

(c) Ponctuation 'exotique' : De même, les signes de ponctuation tels que les parenthèses ou les guillemets conduisent à des différences d'annotation.

(d) Ponctuation implicite : Le rapport des annotateurs précise que « Les principaux points litigieux concernent des endroits où une pause non signalée par la ponctuation pourrait éventuellement être marquée ». Que la ponctuation puisse être marquée ou non relève dans ces cas-là de la pure évaluation personnelle de l'annotateur, rendant ces pauses plus difficilement objectivables. Dans une certaine mesure, les désaccords dus aux dialogues sont une forme spécifique de ponctuation implicite.

Ref	T	Localisation	AB	BC	MH	Moy.	É.-t.
(a)	4	« – Ah ! Ça P c'est drôle ... »	4	4	-10	-0.66	6.6
(b)	4	« je ne dessinerai pas mon avion, c'est un dessin beaucoup trop compliqué pour moi »	6	-10	4	0	7.12
(c)	4	« quand il aperçut pour la première fois mon avion (je ne dessinerai pas mon avion »	-10	8	4	0.66	7.72
(d)	0	« J'ai donc dû choisir un autre métier P et j'ai appris à piloter des avions. »	0	6	0	2	6.8

TABLEAU 15.4 – Exemples de désaccords d'annotations. *T* : Texte ; *Moy.* : Moyenne, *É.-t.* : Écart-type.

Approche statistique

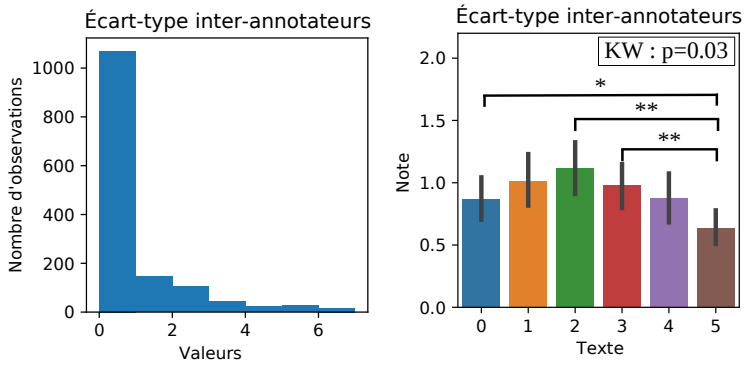

FIGURE 15.5 – **Gauche :** Histogramme des écarts-types interannotateurs. **Droite :** Distribution des écarts-types interannotateurs en fonction des textes (moy. ± é.-t.). *KW :* Test de Kruskal-Wallis ; Tests de Mann-Whitney : * : $p < 0.05$, ** : $p < 0.01$.

D'un point de vue statistique, le désaccord entre annotateurs peut se traduire par l'écart-type interannotateurs. Afin d'avoir une vue générale de ces désaccords, nous avons représenté dans la figure 15.5 un histogramme des écarts-types interannotateurs. Alors que presque 75% des annotations semblent conduire à un consensus solide ($\sigma < 1$), certaines pauses conduisent à des notes plus dispersées, dont l'écart-type d'annotation est parfois supérieur à 6 (cf. tableau 15.4).

Dans un premier temps, nous avons donc cherché quels textes conduisent au meilleur accord interannotateurs. Nous avons donc fait un test statistique de Kruskal-Wallis afin d'objectiver les variations de consensus d'annotation entre les textes, représentées figure 15.5.

Les différences observées sont statistiquement significatives (KW, $H = 12.1, ddof = 5, p = 0.03$). Une analyse post-hoc (tests de Mann-Whitney appliqués à chaque paire de textes) permet de mettre en lumière une différence significative entre le texte n°0 et le texte n°5 ($p = 0.02$), entre le texte n°2 et le texte n°5 ($p = 2.4 \times 10^{-3}$) et entre les textes n°3 et n°5 ($p = 4.5 \times 10^{-3}$). Le texte 5 est ainsi celui conduisant au meilleur consensus entre les trois annotateurs, avec des écarts-types interannotateurs plus faibles que sur les autres textes.

Confrontation des deux approches

En confrontant les deux approches des lieux de désaccords – expérience des annotateurs et approche statistique, un plus grand écart-type interannotateurs moyens sur un texte devrait être la conséquence d'un plus grand nombre de lieux de désaccords dans ce texte. Nous avons donc dénombré et reporté dans le tableau 15.5 les lieux de désaccords objectifiables selon les critères précédemment établis.

Texte	n°0	n°1	n°2	n°3	n°4	n°5
Ponctuation 'exotique'	2	0	2	0	2	4
Virgules	15	16	14	10	14	18
Dialogues	0	3	4	9	8	9
Total désaccord	17	19	20	19	24	31
Total pauses	217	253	248	229	237	252

Tableau 15.5 – Nombre de lieux de désaccords potentiels dans chaque texte.

Ce tableau fait apparaître une contradiction entre la moyenne des écarts-types sur le texte n°5, qui est significativement plus faible que sur les autres textes (figure 15.5) et le nombre de lieux potentiels de désaccord, pour lesquels il en possède le plus (tableau 15.5).

Une première hypothèse pour expliquer cette différence repose sur les pauses implicites, qui n'ont pas été incluses dans le précédent tableau par manque de critère objectif de mesurabilité. Le texte n°2 serait alors celui avec le plus de pauses implicites. Cette catégorie de pauses, qui a été identifiée par les annotateurs comme causant le plus désaccords, est en dehors de la portée de nos analyses statistiques et reste du domaine de l'appréciation subjective de chaque annotateur.

Une deuxième hypothèse, pouvant expliquer la décroissance des désaccords interannotateurs à partir du texte 2, pourrait être un effet d'entraînement. Au fur et à mesure qu'ils gagnent en expérience et en aise, les trois annotateurs mesureraient de manière différente, mais plus précise un même score objectif.

15.3.5 Agrément interannotateurs

Méthode

Afin de déterminer si les différences observées précédemment compromettent l'utilisation de la moyenne des trois annotateurs comme vérité terrain fiable, nous étudions dans cette partie l'agrément interannotateurs de l'annotation des pauses. Pour cela, nous calculons la corrélation intraclasse – *Intraclass Correlation Coefficient* (ICC), une mesure d'agrément permettant de prendre en compte le nombre total d'observations dans chaque classe d'annotation.

L'ICC se décline en 6 mesures différentes suivant l'hypothèse que l'on veut tester (Shrout et Fleiss, 1979). Nous désirons à la fois généraliser les résultats à toute population qui a les mêmes caractéristiques que nos annotateurs (c'est à dire aux orthophonistes de manière

générale) et utiliser la moyenne des annotations comme vérité terrain : nous utilisons donc l'ICC de type 2 dans sa version estimant la fiabilité de la moyenne des trois annotations (notée ICC2-k dans la suite).

Une ICC de 1 représente un accord parfait entre les annotateurs, tandis qu'une ICC de 0 représente un désaccord parfait (Koo et Li, 2016). L'ICC2-k présentée dans cette section a été calculée grâce à la librairie Python `pingouin` (Vallat, 2018).

ICC globale

L'ICC2-k calculée sur chaque texte et pour chaque combinaison d'annotateurs est représentée dans la figure 15.6.

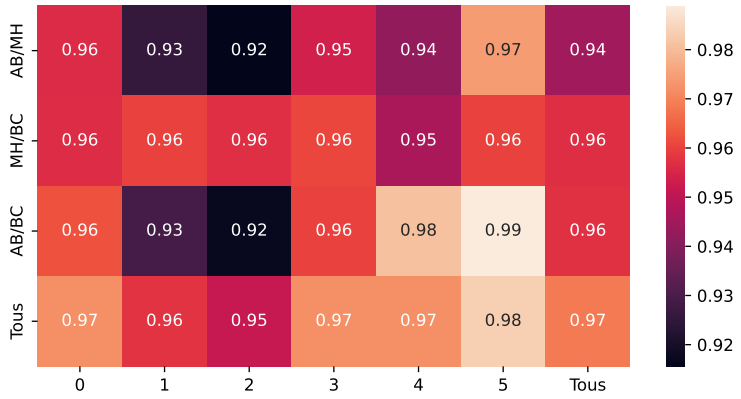

FIGURE 15.6 – ICC2-k en fonction des textes et des combinaisons d'annotateurs.

Toutes les valeurs de cette matrice sont supérieures à 0.9, avec une ICC tous annotateurs et tous textes confondus atteignant 0.97. Cela traduit un effet minoritaire des désaccords entre annotateurs dans les notes moyennes et témoigne d'une excellente fiabilité de l'annotation moyenne (Koo et Li, 2016).

Cette matrice est également cohérente avec les observations faites sur l'écart-type interannotateurs. En effet, les deux minimums sont observés sur le texte n°2, qui est celui ayant le plus grand écart-type interannotateurs moyen. Au contraire, les meilleurs consensus sont observés sur le texte 5, sur tous les annotateurs (ICC = 0.98), et plus particulièrement entre AB et BC (ICC = 0.99), ce qui avait déjà été observé en étudiant l'écart-type interannotateurs.

ICC sur les pauses négatives et positives

Nous proposons également les mêmes analyses exclusivement sur les pauses négatives et les pauses positives respectivement en haut et en bas de la figure 15.7.

Cette analyse permet de conforter nos précédentes observations et d'affiner les origines de celles-ci :

— la session durant laquelle l'accord interannotateurs sur les pauses négatives est le plus faible est la n°3 (ICC = 0.54). Nous avions déjà observé un maximum de l'écart-type interannotateurs (figure 15.5) sur cette session : nous pouvons maintenant affiner la précédente observation et expliquer ce minimum par les désaccords sur l'attribution des notes négatives. Il est cependant à noter que le minimum d'accord sur ces pauses est obtenu entre AB et BC sur le second texte (ICC = 0.34) ;

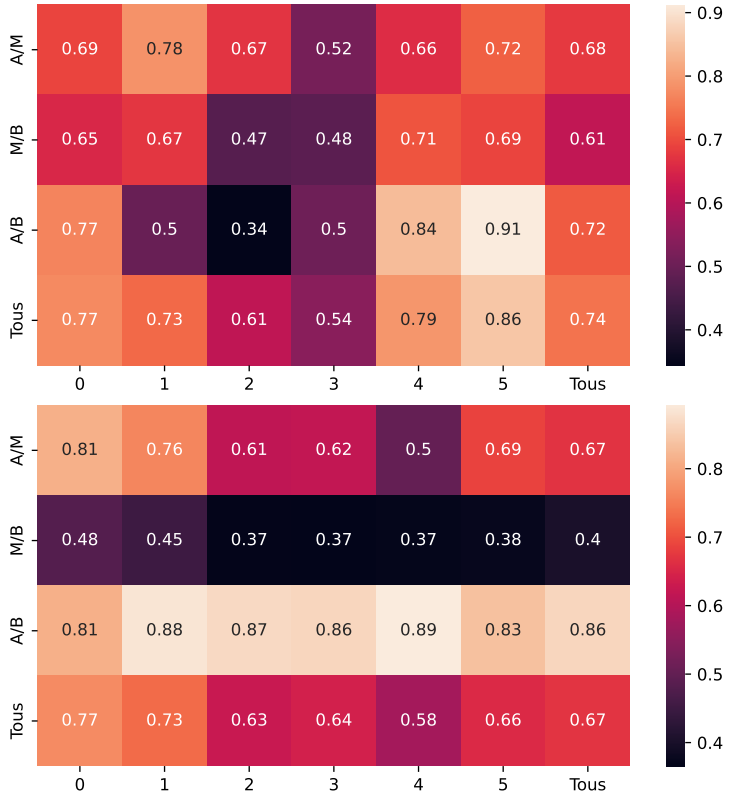

FIGURE 15.7 – ICC2k en fonction des textes et des combinaisons d'annotateurs sur les pauses négatives (haut) et positives (bas).

— les différences de pratique d'annotation précédemment rapportées par MH et BC sur les pauses positives se retrouvent dans la figure 15.7, sur laquelle le minimum des agréments interannotateurs est obtenu entre ces deux annotateurs, quelle que soit la session considérée (ICC ≤ 0.48).

15.3.6 Conclusion

Malgré la présence de lieux de désaccords identifiés, mais pas toujours mesurables de manière objective, les très bonnes valeurs d'accord interannotateurs tous textes et tous annotateurs confondus permettent de valider l'utilisation de la moyenne interannotateurs comme vérité terrain pour la future extraction automatique de marqueurs. De plus, l'ICC2 simple tous textes et annotateurs confondus (non représenté dans les figures précédentes) atteint une valeur plus faible que celle évaluant la robustesse de la moyenne (resp. 0.91 et 0.97) : l'utilisation de l'annotation moyenne au lieu de celle d'un seul annotateur conduit à une vérité terrain plus fiable concernant la naturalité des emplacements des pauses de lecture.

15.4 Analyse des profils de lecteurs

15.4.1 Objectif

En combinant l'estimation automatique des emplacements et des durées des pauses de lecture avec les précédentes annotations, il est possible de calculer, pour chaque enregistrement,

un score évaluant à quel point les pauses marquées par le lecteur sont naturelles.

Au contraire des précédents chapitres pour lesquels nous utilisons des techniques d'apprentissage automatique pour la validation de nos marqueurs, nous proposons dans cette partie une analyse différente, reposant sur des profils de lecteurs. Cette approche permet de ne pas faire d'hypothèse *a priori* sur le lien entre somnolence et pauses de lecture. En nous basant sur les descripteurs de pause de lecture introduits dans la prochaine section, nous créons des profils de lecteurs, dont nous analysons rétrospectivement les caractéristiques médicales afin de mettre éventuellement au jour une concordance entre profils de lecture et profils médicaux.

Conformément à la section précédente, le score utilisé comme vérité terrain pour le calcul des descripteurs est la moyenne des annotations faites par AB, BC et MH. Cette analyse est effectuée sur le corpus TILE-93 présenté dans le chapitre 7.

15.4.2 Marqueurs calculés

À partir de la durée et de l'emplacement des pauses de lecture, nous calculons 4 ensembles de descripteurs pour chaque fichier audio. Chacun d'eux est calculé à la fois sur toutes les pauses (notées *toutes*), sur les pauses notées avec un score positif (notées +) et leur équivalent négatif (notées -). Ces marqueurs sont les suivants :

— N : le nombre de pauses ;
— D : la durée moyenne des pauses ;
— S : le score moyen des pauses ;
— WS : le score moyen des pauses pondéré par leur durée.

Par exemple, en notant d_i et s_i la durée et le score de la i-ème pause, nous avons :

$$\mathrm{WS}^+ = \frac{1}{\mathrm{N}^+}\frac{1}{d^+_{tot}}\sum_{i\in+} d_i s_i \quad \text{avec : } d^+_{tot} = \sum_{i\in+} d_i$$

Un exemple de pauses positives et négatives avec leur durée, score, et score pondéré est proposé dans la figure 15.8.

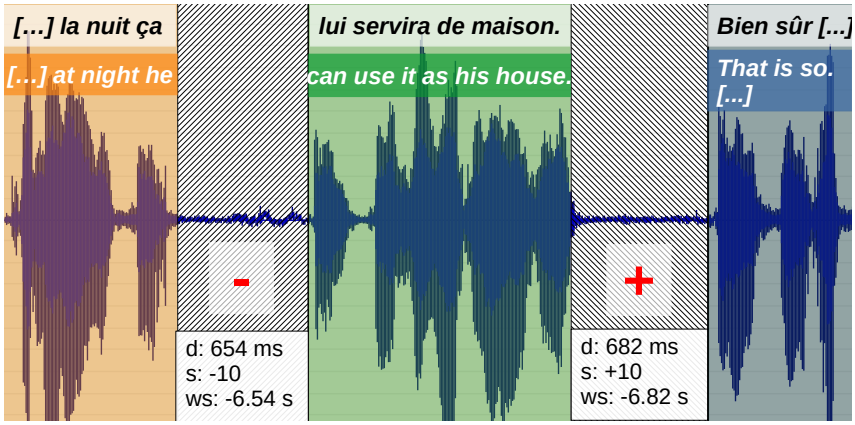

FIGURE 15.8 – Exemple de pauses incorrectement et correctement placées, avec leur score, leur durée et le score pondéré correspondant.

Lors des études suivantes des caractéristiques des patients, ces marqueurs sont moyennés sur les cinq enregistrements effectués au cours du TILE.

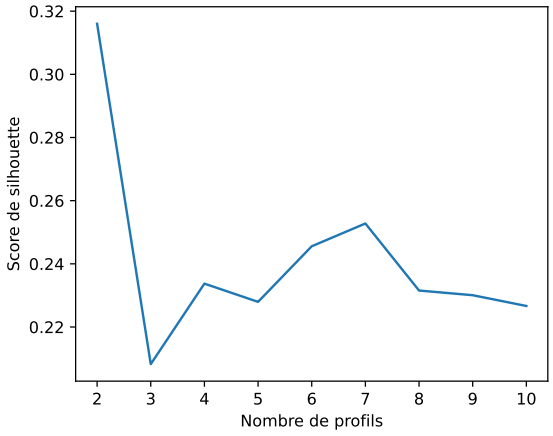

FIGURE 15.9 – Score de silhouette en fonction du nombre de profils lors de l'algorithme *KMeans*.

15.4.3 Profils de lecteur

15.4.4 Méthode

Afin de construire des profils de lecteurs, nous appliquons un algorithme de *K-Means* sur les marqueurs moyens normalisés (centré et réduits) des locuteurs. Le nombre de groupes $n < 10$ est choisi de façon à maximiser le score de silhouette total.

Ce dernier est défini comme la somme sur tous les échantillons du score de silhouette, défini par l'équation suivante :

$$s_i = \frac{b_i - a_i}{\max(a_i, b_i)}$$

avec a_i la distance intragroupe moyenne et b_i la distance moyenne avec le groupe le plus proche auquel n'appartient pas l'échantillon i.

Le score de silhouette total pour différents groupes est présenté dans la figure 15.9. Le score est ainsi maximal pour $k = 7$ (score de silhouette total : 0.253) : les patients du corpus sont répartis selon 7 profils de comportements concernant les pauses de lectures.

15.4.5 Analyse en composantes principales

Une première étape pour analyser les profils obtenus est d'appliquer une analyse en composantes principales (ACP) avec $n = 2$ composantes sur les marqueurs de pauses de lecture (centrés et réduits). Les coefficients ainsi obtenus sont reportés dans le tableau 15.6.

Ne conserver que deux dimensions dans l'ACP permet de répliquer 75.5% de la variance originale et conserve un bon contraste entre les sept profils précédemment identifiés (score de silhouette total : 0.246). Les sept profils en fonction de leur projection sur les deux dimensions de l'ACP sont tracés dans la figure 15.10.

Ainsi, les deux principales dimensions sur cet ensemble de marqueurs sont dirigées d'une part par le nombre total de pauses et leur durée, indépendamment de leur justesse (Dim. 1), et d'autre part par la différence d'occurrence et de durée entre les pauses positives et négatives (Dim. 2).

Ces deux dimensions sont anti-corrélées avec S^- et WS^-, et correspondent à deux comportements de mauvaise lecture : la dimension Dim. 1 correspond à un nombre important et une

durée excessive d'arrêt, tandis que Dim. 2 mesure la propension à s'arrêter à des emplacements peu naturels.

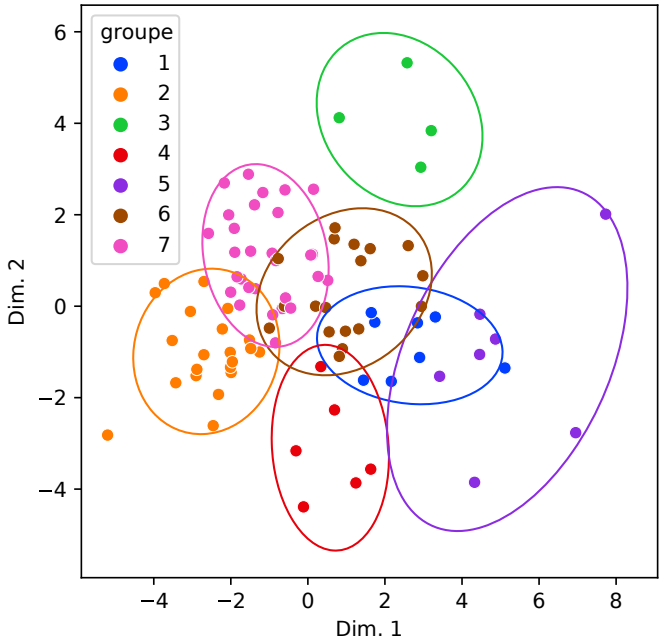

FIGURE 15.10 – Représentation graphique des profils obtenus par l'algorithme *KMeans* en fonction des deux dimensions de l'ACP. Les ellipses ont été tracées en utilisant des rayons égaux à 2σ.

	+				-				all			
	N	D	S	WS	N	D	S	WS	N	D	S	WS
Dim. 1	.28	.36	.20	.34	.25	.33	-.27	-.27	.34	.36	-.20	.16
Dim. 2	-.23	-.20	-.06	-.23	.36	.21	-.25	-.25	-.04	-.20	-.44	-.49

TABLEAU 15.6 – Coefficients de l'ACP pour les différents descripteurs liés aux pauses.

15.4.6 Analyses statistiques

Méthode

Afin de mettre au jour la relation entre les caractéristiques physiques et médicales des patients, et leur comportement de lecture, nous procédons en trois étapes :

1. **Déviation à la moyenne** : pour chacun des sept profils de lecteur et pour chaque descripteur, nous rapportons dans le tableau 15.7 la distance entre le centre du cluster et la valeur moyenne de ce descripteur sur le corpus entier. Nous rapportons également au bas de ce tableau les mêmes métriques calculées sur les informations médicales des patients. Ces résultats servent de support central pour la discussion de la section 15.4.7.
2. **ANOVA** : afin d'identifier les facteurs influençant le plus les pauses lors de la lecture, nous calculons une ANOVA multivariée à mesures répétées, indépendamment des profils. De cette façon, nous essayons d'expliquer les variations intra et inter locuteurs des descripteurs avec les caractéristiques médicales des patients. Afin de gagner en généralité sans perdre trop de puissance statistique, nous calculons uniquement les effets des

variables simples, et des interactions de premier et second ordre entre l'ESS, le TILE moyen et les cofacteurs des locuteurs. Le résultat de ce test statistique est présenté dans le tableau 15.8.

3. **Regression** : Nous avons également reporté dans le tableau 15.8 les corrélations univariées entre les co-facteurs et les descripteurs des pauses.

Descripteur		m (σ)	**n°1** *n=8*	**n°2** *n=21*	**n°3** *n=4*	**n°4** *n=6*	**n°5** *n=7*	**n°6** *n=18*	**n°7** *n=29*
+	N	23.3 (5.4)	=	- -	=	+++	+++	++	- -
	D	615.5 ms (143.0)	+++	- -	=	++	+++	=	- -
	S	8.0 (0.3)	=	+++	- - -	-	- - -	- -	- -
	WS	5.0 (1.1)	+++	- -	-	++	+++	=	- -
-	N	3.5 (2.7)	=	- -	+++	- -	+++	=	=
	D	480.3 ms (140.3)	+++	- - -	+++	- -	+++	+	=
	S	-7.3 (1.7)	- -	+++	- -	+++	=	-	- -
	WS	-4.0 (1.3)	- - -	+++	- -	+++	- -	=	=
Toutes	N	26.8 (6.4)	=	- -	+++	++	+++	++	- -
	D	605.3 ms (132.7)	+++	- -	=	++	+++	=	- -
	S	6.0 (1.3)	=	++	- - -	++	- -	=	=
	WS	3.8 (1.1)	+++	=	- - -	+++	+++	=	- -
Sexe %F		62.4	25.0	85.7	75.0	50.0	14.3	44.4	79.3
Age		36.6 (14.5)	++	-	- -	++	++	+	=
IMC		23.9 (5.1)	=	=	-	=	=	=	=
Cou		37.8 (4.4)	+	-	- -	=	++	+	=
Édu.		5.5 (2.6)	- -	+	- -	+	=	-	=
HAD-A		8.5 (4.2)	=	=	+++	=	-	=	=
HAD-D		6.7 (3.8)	- - -	-	+	+	=	+	=
ESS		14.6 (4.7)	-	=	-	++	- -	=	=
TILE moy.		11.6 (4.6)	- -	=	=	=	=	=	=

TABLEAU 15.7 – Profils des lecteurs en fonction de leurs pauses. m : valeur moyenne sur le corpus entier, σ écart-type sur le corpus entier.
Notations en fonction du centre du cluster m_c. = : $m_c \in [m - 0.25\sigma; m + 0.25\sigma]$; - : $m_c \in [m - 0.5\sigma; m - 0.25\sigma]$; - - : $m_c \in [m - \sigma; m - 0.5\sigma]$; - - - : $m_c < m - \sigma$. +, ++ et +++ sont les notations équivalentes pour $m_c > m$.

Tableau 15.8 – Analyse statistique des marqueurs proposés au regard des autres caractéristiques des patients. Haut : significativité des facteurs inclus dans l'ANOVA multivariée à mesures répétées conçues pour expliquer les variations intra- et interlocuteurs de chaque descripteur. Bas : Corrélation entre les descripteurs et les caractéristiques des patients. Les corrélations à l'échelle des locuteurs sont calculées avec la moyenne des marqueurs au cours des itérations du TILE. * : $p < 0.05$, ** : $p < 0.01$, *** : $p < 0.001$.

Facteur	**Nombre**			**Durée (ms.)**			**Score**			**Score pondéré**		
	+	-	toutes	+	-	toutes	+	-	toutes	+	-	toutes
Moy.	23.3	3.5	26.8	615.5	480.3	605	8.0	-7.3	6.0	5.0	-4.0	3.8
É.-t.	6.3	3.4	7.5	156.5	241.9	144.7	.53	3.2	1.7	1.3	2.33	1.4
ANOVA												
Sexe	**		**	**		**	*			**		*
Âge				***		***				***		***
IMC					**			*			**	
Cou	*		*	*	**	*	*					
Socio-D.	*		*				*					
HAD-A	*								*			
ESS					*						*	
TILE : ESS : IMC		*							*			
TILE : ESS : Socio-Dem								**			*	
Session	***	***	***	*			***	***	***	***	***	***
KSS				td	**			*		*	**	*
KSS : TILE it							td					
Correlations : ρ de Spearman (sig.)												
Age	.21 (*)			.44 (***)		.42 (***)				.45 (***)		.40 (***)
Cou	.3 (**)	.26 (*)	.29 (**)	.32 (**)	.33 (**)	.33 (**)	-.32 (**)		-.23 (*)	.28 (**)	-.21 (*)	
Socio-Dem.		-.29 (**)	-.28 (**)	-.21 (*)	-.21 (*)		.30 (**)		.27 (**)			
KSS		.13 (**)		.12 (**)	.09 (*)	.13 (**)	-.09 (*)		-.15 (**)	.09 (*)		

Résultats

ANOVA Les variations intralocuteurs de tous les marqueurs exceptés les durées sont fortement influencées par les effets de session ($p < 0.001$) : les variations des descripteurs au cours d'une session sont soit influencées par le moment de la journée durant lequel l'enregistrement a été fait, soit par le texte qui change à chaque itération du TILE. Une partie de la variation de D^+ et D^-, N^+, et de WS^+, WS^- et WS^{toutes} peut être attribuée à la somnolence subjective. Puisque les latences d'endormissement au TILE prises de manière individuelles n'influencent l'évolution d'aucun marqueur, celles-ci n'ont pas été rapportées dans le tableau. Enfin, l'interaction commune de la KSS et de la latence d'endormissement au TILE ont un faible effet sur le score des pauses positives.

Alors que le sexe influence toutes les caractéristiques liées aux pauses positives, l'âge affecte seulement D^+ et WS^+. Au contraire, l'IMC semble n'affecter que D^- et S^-. Le tour de cou interfère avec N^+ et S^+, et D^{toutes}. Le niveau sociodémographique a le même effet sur les pauses positives, et le niveau d'anxiété a une influence sur N^+ et S^{toutes}. Le niveau de dépression n'a aucune influence sur les marqueurs liés aux pauses, le HAD-D n'a donc pas été reporté dans ce tableau. Concernant la somnolence, D^- est influencé par l'ESS, et N^- et S^{toutes} sont influencées par l'interaction conjointe du TILE, de l'ESS et de l'IMC. Enfin, l'interaction conjointe du TILE, de l'ESS et du niveau sociodémographique interagit avec S^-.

Régression À l'échelle des itérations, la latence d'endormissement ne corrèle avec aucun autre facteur et n'a pas été reportée dans le tableau. Au contraire, la somnolence subjective corrèle faiblement, mais significativement avec S^-, D^+, D^- et D^{toutes}, et WS^+ et anti-corrèle avec S^+ et S^{toutes}.

À l'échelle des locuteurs, seuls l'âge, le tour de cou et le niveau sociodémographique corrèlent significativement avec les caractéristiques de pauses. Ainsi, l'âge corrèle avec N^+, D^+, WS^+, D^{toutes} et WS^{toutes}. Le niveau sociodémographique et le tour de cou ont des effets antagonistes sur D^-, D^{toutes}, N^+ et N^-, S^+, et S^{toutes}. Le tour de cou est également corrélé à N^+, WS^+ et anti-corrélé à WS^-.

15.4.7 Discussion

Dans cette section, nous proposons une interprétation des différents profils obtenus au regard des résultats précédents.

Patients avec une faible qualité de lecture

Profil n°1 Les locuteurs appartenant au profil n°1 ($n = 8$) présentent deux comportements de lecture anormaux : non seulement ils allongent leurs pauses à la fois là où c'est naturel et là où ça ne l'est pas (Dim. 1 de l'ACP), mais en plus les emplacements de leurs pauses négatives sont pires que la moyenne (Dim. 2 de l'ACP). Concernant les pauses positives, ces patients sont plus âgés et ont un tour de cou plus élevé que la moyenne de notre population, ce qui induit de plus grandes valeurs de D^+ et WS^+. Concernant les comportements de lecture négatifs, les plus grandes valeurs de D^- peuvent être dues à la corrélation entre ce marqueur et le tour de cou, tandis que les plus faibles valeurs de S^- et WS^- peuvent être expliquées par l'interaction du TILE, de l'ESS et du niveau sociodémographique, qui sont plus faibles que la moyenne.

Profil n°3 Les locuteurs du profil n°3 ($n = 4$) ont des pauses excessivement mal placées, ce qui se traduit par des pauses négatives plus nombreuses et qui durent plus longtemps que le patient moyen (Dim. 2 de l'ACP). L'augmentation de N^- et D^- et la diminution de S^+ peuvent être expliquées par leur niveau sociodémographique plus faible, qui n'est pas compensé par l'effet antagoniste du tour de cou, trop faible dans ce groupe. De plus, ces patients sont caractérisés par un plus haut niveau d'anxiété, qui peut expliquer les plus faibles valeurs de S^{toutes}. Similairement aux locuteurs du profil n°1, les plus faibles valeurs de S^- et WS^- sont expliquées par l'influence conjointe du TILE, de l'ESS et du niveau d'éducation.

Profil n°5 Les locuteurs appartenant au profil n°5 ($n = 7$) représentent une troisième catégorie de locuteurs ayant une mauvaise qualité de lecture liée aux pauses. D'une part, ils produisent plus de pauses correctement placées, mais avec un score en moyenne plus faible (Dim. 1 de l'ACP) : cela semble indiquer que le lecteur ou la lectrice fait des efforts de nuances, s'arrêtant aux pauses implicites du texte en plus de celles indiquées par la ponctuation. Ce comportement semble lié à l'âge des locuteurs et locutrices de ce profil, qui est plus élevé que la moyenne. D'un autre côté, leurs pauses (à la fois positives et négatives) sont plus longues que la moyenne, et ils font plus des pauses négatives sans particulièrement les placer aux endroits les moins appropriés, contrairement aux profils n°1 et n°3. Ce comportement peut être expliqué par la plus grande circonférence du cou des locuteurs pour S^+, N^- et D^-, ou par une ESS plus faible pour D^-.

Patients avec une qualité de lecture au-dessus de la moyenne

Les profils identifiés ne correspondent pas qu'à des lecteurs avec une faible qualité de lecture.

Profil n°2 En effet, les patients du profil n°2 ($n = 21$) font moins de pauses correctement placées que la moyenne, mais elles sont en moyenne mieux placées (Dim. 1 de l'ACP). En particulier, ils possèdent une caractéristique commune avec les lecteurs du profil n°4, à savoir un faible nombre de pauses négatives, qui sont plus courtes et qui ont en moyenne des scores et scores pondérés élevés (Dim. 2 de l'ACP). Cela traduit un plus faible nombre d'erreurs de pauses, qui sont corrigées plus rapidement. Tandis que la diminution de N^+ peut être expliquée par des circonférences de cou plus faibles, les diminutions de D^+ et WS^+ sont liées à un plus faible âge moyen. De plus, l'augmentation de S^+ et la diminution de N^- et D^- peuvent être expliquées par des tours de cou plus faibles et un plus grand niveau sociodémographique de ces lecteurs ou lectrices. Les plus grandes valeurs de S^- et WS^- ne peuvent pas être expliquées par nos analyses.

Profil n°4 Les patients du profil n°4 ($n = 4$) partagent le comportement de correction rapide des pauses mal placées observé dans le profil n°2. Concernant ce profil, l'âge, plus faible, explique les plus grandes valeurs de N^+ et D^+, et le niveau sociodémographique pourrait être responsable des valeurs plus faibles observées pour N^- et D^-. Ce dernier comportement peut également être expliqué par l'ESS, qui est plus élevée dans ce profil. Enfin, les plus grandes valeurs de S^- et WS^- peuvent être expliquées par l'interaction conjointe du TILE, de l'ESS et de l'IMC.

Autres profils

Profil n°6 Les lecteurs et lectrices du profil n°6 ($n = 18$) ont tendance à s'arrêter un peu plus souvent là où c'est naturel de le faire, mais ce comportement de lecture ne s'écarte pas significativement de la moyenne. Ces faibles variations pourraient être expliquées par les variations correspondantes de l'âge, du tour de cou, et du niveau sociodémographique de la même façon que pour les autres profils.

Profil n°7 Enfin, les patientes et patients correspondant au profil n°7 ($n = 29$) s'arrêtent un peu moins aux endroits où cela est naturel de le faire, et ont tendance à moins bien placer leurs pauses ; mais de manière similaire au profil précédent, ces tendances ne sont pas significativement différentes de la population moyenne de notre corpus.

15.5 Conclusion

En conclusion, nous avons établi sept profils à partir des pauses de lecture de 93 patients atteints d'hypersomnie. Ces profils sont orientés d'une part par le nombre total et la durée totale des pauses, et d'autre part par le nombre et la durée des pauses négatives. Ces comportements de lecture sont liés aux caractéristiques des patients, avec une influence prédominante de l'âge, de la taille du cou et du niveau sociodémographique, mais aussi une influence de la somnolence diurne excessive par l'influence conjointe de la somnolence objective et subjective et du niveau sociodémographique.

Au niveau de la session, ces caractéristiques ne sont pas corrélées avec la latence d'endormissement au TILE et faiblement avec la KSS, l'effet de session masquant toutes les autres différences : ces descripteurs sont mieux adaptés à l'estimation des traits du locuteur que de son état.

Conclusion de la partie

Nous avons, dans cette partie, introduit et validé quatre groupes de descripteurs vocaux basés sur des enregistrements de sujets en train de lire un texte à voix haute.

Marqueurs acoustiques

Nous avons tout d'abord conçu des descripteurs de la qualité acoustique de la voix, moins nombreux et plus simples que ceux proposés dans l'état de l'art, dont l'interprétabilité a été validée avec des médecins du sommeil. S'ils permettent d'atteindre des performances du niveau de l'état de l'art sur un sous-corpus du *Sleepy Language Corpus* composé uniquement de tâches de lecture (77.6% d'UAR), ces marqueurs ne permettent pas une bonne classification ni de la somnolence objective ni de la somnolence subjective sur le corpus TILE (UAR $< 60\%$).

Erreurs de lecture

En conséquence, nous avons conçu de nouveaux marqueurs, mesurant l'influence de la somnolence sur les capacités cognitives, et plus particulièrement les capacités de lecture à voix haute : les erreurs de lecture.

Nous avons annoté la base TILE et étudié quatre sous-catégories d'erreurs de lecture : les achoppements, les paralexies, les additions et les délétions.

Ces erreurs, associées à un classificateur (SVM) permettent d'atteindre des scores de classification suffisants pour leur utilisation clinique dans la cadre de la détection d'une latence d'endormissement pathologique au TILE (82.6% d'UAR). Cependant, elles nécessitent d'être annotées à la main, processus long, fastidieux et nécessitant une formation spécifique.

Erreurs des systèmes de transcription automatique

Nous avons donc voulu automatiser l'extraction de ces erreurs de lecture en utilisant des systèmes de transcription automatique (STA). Pour cela, nous avons étudié quatre types d'erreurs (substitutions, délétions, insertions, nombre d'unités correctement décodées) faites par sept systèmes de STA bout-en-bout.

Associées à un schéma de classification assurant la *sensibilité* et la *spécificité* des marqueurs sélectionnés par le classifieur, ces erreurs permettent la classification correcte d'une latence d'endormissement pathologique au TILE (73.2% d'UAR) et, en combinaison avec les précédents marqueurs, de la somnolence diurne excessive sévère (ESS > 15, 74.2% d'UAR). Ainsi, les performances des systèmes automatiques nous permettent d'écarter les erreurs de lecture annotées manuellement.

Durée et *naturalité* des pauses de lecture

Enfin, la dernière catégorie de marqueurs étudiés est les durées, le nombre et les emplacements des pauses faites lors de la lecture. Nous avons d'abord proposé et validé un système permettant d'estimer automatiquement l'emplacement précis des pauses extraites dans le texte lu à partir des enregistrements audio.

Nous avons ensuite demandé à trois élèves en M1 d'orthophonie d'annoter ces textes avec, pour chaque paire de mots, la *naturalité* de la pause entre ceux-ci. Nous avons validé la fiabilité de ces annotations, notamment dans le cadre d'une utilisation en tant que vérité terrain pour estimer la naturalité des pauses faites par les locuteurs.

Enfin, à partir de ces annotations, nous avons extrait automatiquement des marqueurs de durée et de *naturalité* des pauses faites par les lecteurs. Ces marqueurs nous ont permis d'identifier sept profils de lecteurs dans le corpus TILE, que nous avons associés aux caractéristiques (âge, niveau d'éducation, IMC, somnolence au long cours) de ces locuteurs.

Prochaine partie Dans la prochaine partie, nous proposons des systèmes de classification de différents symptômes et différents syndromes liés à la somnolence au long cours, basés sur les marqueurs élaborés dans cette partie et sur des techniques d'apprentissage automatique.

Bibliographie de la partie

Albuquerque, L., Valente, A. R. S., Teixeira, A., Figueiredo, D., Sa-Couto, P., et Oliveira, C. (**2021**). "Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan," PLOS ONE **16**(4), e0248842, doi: 10.1371/journal.pone.0248842.

Aldrich, M. S., Chervin, R. D., et Malow, B. A. (**1997**). "Value of the multiple sleep latency test (MSLT) for the diagnosis of narcolepsy," Sleep **20**(8), 620–629, doi: 10.1093/sleep/20.8.620.

Amiriparian, S., Winokurow, P., Karas, V., Ottl, S., Gerczuk, M., et Schuller, B. (**2020**). "Unsupervised Representation Learning with Attention and Sequence to Sequence Autoencoders to Predict Sleepiness From Speech," dans *MM '20 : The 28th ACM International Conference on Multimedia*, pp. 11–17, doi: 10.1145/3423327.3423670.

Arand, D., Bonnet, M., Hurwitz, T., Mitler, M., Rosa, R., et Sangal, R. B. (**2005**). "The Clinical Use of the MSLT and MWT," SLEEP **28**(1), 123–144, doi: 10.1093/sleep/28.1.123.

Balagopalan, A., Eyre, B., Robin, J., Rudzicz, F., et Novikova, J. (**2021**). "Comparing Pre-trained and Feature-Based Models for Prediction of Alzheimer's Disease Based on Speech," Frontiers in Aging Neuroscience **13**, doi: 10.3389/fnagi.2021.635945.

Bose, A., Dash, N. S., Ahmed, S., Dutta, M., Dutt, A., Nandi, R., Cheng, Y., et D Mello, T. M. (**2021**). "Connected Speech Characteristics of Bengali Speakers With Alzheimer's Disease : Evidence for Language-Specific Diagnostic Markers," Frontiers in Aging Neuroscience **13**, 707628, doi: 10.3389/fnagi.2021.707628.

Boyer, F. (**2021**). "Reconnaissance automatique de parole et intégration dans un système de compréhension du langage parlé," Thèse de doctorat, Université de Bordeaux.

Boyer, F., et Rouas, J.-L. (**2019**). "End-to-End Speech Recognition : A review for the French Language," arXiv : 1910.08502 .

Boyer, S., El-Yagoubi, R., Tiberge, M., Ruiz, R., et Daurat, A. (**2016**). "Paramètres Acoustiques de la Voix et Privation de Sommeil," dans *CFA/VISHNO*.

Bozkurt, E., Erzin, E., Erdem, \. E., et Erdem, A. T. (**2011**). "RANSAC-based training data selection for speaker state recognition," dans *Interspeech 2011*, pp. 3293–3296, doi: 10.21437/Interspeech.2011-811.

Brin, F., Courrier, C., Lederle, E., et Masy, V. (**2018**). *Dictionnaire d'orthophonie - 4ème édition*, orthoedition éd.

Chawla, N. V., Bowyer, K. W., Hall, L. O., et Kegelmeyer, W. P. (**2002**). "SMOTE : Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research **16**, 321–357, doi: 10.1613/jair.953.

Chen, Y., Dinkel, H., Wu, M., et Yu, K. (**2020**). "Voice Activity Detection in the Wild via Weakly Supervised Sound Event Detection," dans *Interspeech 2020*, pp. 3665–3669, doi: 10.21437/Interspeech.2020-995.

Clarke, N., Barrick, T., et Garrard, P. (**2021**). "A Comparison of Connected Speech Tasks for Detecting Early Alzheimer's Disease and Mild Cognitive Impairment Using Natural Language Processing and Machine Learning," Frontiers in Computer Science **3**, doi: 10.3389/fcomp.2021.634360.

Degottex, G., Kane, J., Drugman, T., Raitio, T., et Scherer, S. (**2014**). "COVAREP — A collaborative voice analysis repository for speech technologies," dans *ICASSP 2014*, pp. 960–964, doi: 10.1109/ICASSP.2014.6853739.

DeJong, S. (**1993**). "SIMPLS : An alternative approach to partial least squares regression," Chemometrics and Intelligent Laboratory Systems **18**(3), 251–263.

Demiroglu, C., Besirli, A., Ozkanca, Y., et Çelik, S. (**2020**). "Depression level assessment from multi-lingual conversational speech data using acoustic and text features," EURASIP Journal on Audio, Speech, and Music Processing **2020**(1), 17, doi: 10.1186/s13636-020-00182-4.

Dhupati, L. S., Kar, S., Rajaguru, A., et Routray, A. (**2010**). "A novel drowiness detection scheme based on speech analysis with validation using simultaneous EEG recordings," dans *IEEE - Int. CASE*, pp. 917–921.

Egas-Lopez, J. V., et Gosztolya, G. (**2021**). "Deep Neural Network Embeddings for the Estimation of the Degree of Sleepiness," dans *ICASSP 2021*, Toronto, ON, Canada, pp. 7288–7292, doi: 10.1109/ICASSP39728.2021.9413589.

Elsner, D., Langer, S., Ritz, F., Mueller, R., et Illium, S. (**2019**). "Deep Neural Baselines for Computational Paralinguistics," dans *Interspeech 2019*.

Eyben, F., et Schuller, B. (**2015**). "Opensmile," ACM SIGMultimedia Records **6**, 4–13.

Freitag, M., Amiriparian, S., Pugachevskiy, S., Cummins, N., et Schuller, B. (**2017**). "audeep : Unsupervised learning of representations from audio with deep recurrent neural networks," The Journal of Machine Learning Research **18**(1), 6340–6344.

Fritsch, J., Dubagunta, S. P., et Magimai.-Doss, M. (**2020**). "Estimating the Degree of Sleepiness by Integrating Articulatory Feature Knowledge in Raw Waveform Based CNNS," dans *ICASSP 2020*, Barcelona, Spain, pp. 6534–6538, doi: 10.1109/ICASSP40776.2020.9053351.

Gajšek, R., Dobrišek, S., et Mihelič, F. (**2011**). "University of Ljubljana system for interspeech 2011 speaker state challenge," dans *Interspeech 2011*, pp. 3297–3300, doi: 10.21437/Interspeech.2011-812.

Galliano, S., Gravier, G., et Chaubard, L. (**2009**). "The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts," dans *Interspeech 2009*, pp. 2583–2586.

Gonzalez-Atienza, M., Peinado, A. M., et Gonzalez-Lopez, J. A. (**2021**). "An Automatic System for Dementia Detection using Acoustic and Linguistic Features," dans *IberSPEECH 2021*, pp. 265–269, doi: 10.21437/IberSPEECH.2021-56.

Google (**2021**). "WebRTC" .

Gosztolya, G. (**2019**). "Using Fisher Vector and Bag-of-Audio-Words Representations to Identify Styrian Dialects, Sleepiness, Baby & Orca Sounds," dans *Interspeech 2019*, pp. 2413–2417, doi: 10.21437/Interspeech.2019-1726.

Greeley, H. P., Friets, E., Wilson, J. P., Raghavan, S., Picone, J., et Berg, J. (**2006**). "Detecting Fatigue From Voice Using Speech Recognition," dans *IEEE International Symposium on Signal Processing and Information Technology*, pp. 567–571.

Gupta, V., Kenny, P., Ouellet, P., et Stafylakis, T. (**2014**). "I-vector-based speaker adaptation of deep neural networks for French broadcast audio transcription," dans *ICASSP 2014*, pp. 6334–6338.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., et Ng, A. Y. (**2014**). "Deep Speech : Scaling up end-to-end speech recognition," arXiv :1412.5567 [cs] .

Haulcy, R., et Glass, J. (**2021**). "Classifying Alzheimer's Disease Using Audio and Text-Based Representations of Speech," Frontiers in Psychology **11**, doi: 10.3389/fpsyg.2020.624137.

Hermansky, H., et Morgan, N. (**1994**). "RASTA processing of speech," IEEE Transactions on Speech and Audio Processing **2**(4), 578–589, doi: 10.1109/89.326616.

Hillenbrand, J., Cleveland, R. A., et Erickson, R. L. (**1994**). "Acoustic Correlates of Breathy Vocal Quality," Journal of Speech, Language, and Hearing Research **37**(4), 769–778, doi: 10.1044/jshr.3704.769.

Huang, D.-Y., Tsao, Y., Chiori, H., et Kashioka, H. (**2011**). "Feature Normalization and Selection for Robust Speaker State Recognition," dans *IEEE - International Conference on Speech Database and Assessments*, doi: 10.1109/ICSDA.2011.6085988.

Huang, D.-Y., Zhang, Z., et Ge, S. S. (**2014**). "Speaker State Classification Based on Fusion of Asymmetric Simple Partial Least Squares (SIMPLS) and Support Vector Machines," Comput. Speech Lang. **28**(2), 392–419.

Kim, S., Kwon, N., O'Connell, H., Fisk, N., Ferguson, S., et Bartlett, M. (**2020**). ""How are you ?" Estimation of anxiety, sleep quality, and mood using computational voice analysis," dans *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, Montreal, QC, Canada, pp. 5369–5373, doi: 10.1109/EMBC44109.2020.9175788.

Koo, T. K., et Li, M. Y. (**2016**). "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," Journal of Chiropractic Medicine **15**(2), 155–163, doi: 10.1016/j.jcm.2016.02.012.

Krajewski, J., Batliner, A., et Golz, M. (**2009**). "Acoustic sleepiness detection : Framework and validation of a speech-adapted pattern recognition approach," Behavior Research Methods **41**(3), 795–804.

Liu, S., Keren, G., Parada-Cabaleiro, E., et Schuller, B. (**2021**). "N-HANS : A neural network-based toolkit for in-the-wild audio enhancement," Multimedia Tools and Applications doi: 10.1007/s11042-021-11080-y.

Martin, V. P., Rouas, J.-L., et Philip, P. (**2021**). "Étude des erreurs de transcription automatique pour la détection de la somnolence à long terme de patients hypersomniaques," dans *9ème Journées de Phonétique Clinique [poster]*, Toulouse (virtuel).

Martin, V. P., Rouas, J.-L., Thivel, P., et Krajewski, J. (**2019**). "Sleepiness detection on read speech using simple features," dans *10th Conference on Speech Technology and Human-Computer Dialogue*, Timisoara, Romania, doi: 10.1109/SPED.2019.8906577.

McGlinchey, E. L., Talbot, L. S., Chang, K.-h., Kaplan, K. A., Dahl, R. E., et Harvey, A. G. (**2011**). "The Effect of Sleep Deprivation on Vocal Expression of Emotion in Adolescents and Adults," Sleep **34**, 1233–1241.

Montacié, C., et Caraty, M.-J. (**2011**). "Combining multiple phoneme-based classifiers with audio feature-based classifier for the detection of alcohol intoxication," dans *Interspeech 2011*, pp. 3205–3208, doi: 10.21437/Interspeech.2011-802.

Muza, R., Lykouras, D., et Rees, K. (**2016**). "The utility of a 5th nap in multiple sleep latency test," Journal of Thoracic Disease **8**(2), 282–286, doi: 10.3978/j.issn.2072-1439.2015.12.66.

Noffs, G., Boonstra, F. M. C., Perera, T., Kolbe, S. C., Stankovich, J., Butzkueven, H., Evans, A., Vogel, A. P., et van der Walt, A. (**2020**). "Acoustic Speech Analytics Are Predictive of Cerebellar Dysfunction in Multiple Sclerosis," The Cerebellum **19**(5), 691–700, doi: 10.1007/s12311-020-01151-5.

Nwe, T. L., Li, H., et Minghui, D. (**2006**). "Analysis and Detection of Speech under Sleep Deprivation," dans *Interspeech 2006*.

Oppenheim, A., et Schafer, R. (**2004**). "Dsp history - From frequency to quefrency : a history of the cepstrum," IEEE Signal Processing Magazine **21**(5), 95–106, doi: 10.1109/MSP.2004.1328092.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., et Duchesnay, E. (**2011**). "Scikit-learn : Machine Learning in Python," Journal of Machine Learning Research **12**, 2825–2830.

Pellegrino, F., et Andre-Obrecht, R. (**2000**). "Automatic language identification : an alternative approach to phonetic modelling," Signal Processing **80**(7), 1231–1244.

Philip, P., Sagaspe, P., Taillard, J., Chaumet, G., Bayon, V., Coste, O., Bioulac, B., et Guilleminault, C. (**2008**). "Maintenance of Wakefulness Test, obstructive sleep apnea syndrome, and driving risk," Annals of Neurology **64**(4), 410–416, doi: 10.1002/ana.21448.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlıcek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., et Vesely, K. (**2011**). "The Kaldi Speech Recognition Toolkit," dans *ASRU 2011*, pp. 1–4.

Rahman, T., Mariooryad, S., Keshavamurthy, S., Liu, G., Hansen, J. H. L., et Busso, C. (**2011**). "Detecting sleepiness by fusing classifiers trained with novel acoustic features," dans *Interspeech 2011*, pp. 3285–3288, doi: 10.21437/Interspeech.2011-809.

Ravi, V., Park, S. J., Afshan, A., et Alwan, A. (**2019**). "Voice Quality and Between-Frame Entropy for Sleepiness Estimation," dans *Interspeech 2019*, pp. 2408–2412, doi: 10.21437/Interspeech.2019-2988.

Rodríguez, A. N. (**2011**). "An HMM-based approach to the INTERSPEECH 2011 speaker state challenge," dans *Interspeech 2011*, pp. 3289–3292, doi: 10.21437/Interspeech.2011-810.

Romana, A., Bandon, J., Perez, M., Gutierrez, S., Richter, R., Roberts, A., et Provost, E. M. (**2021**). "Automatically Detecting Errors and Disfluencies in Read Speech to Predict Cognitive Impairment in People with Parkinson's Disease," dans *Interspeech 2021*, pp. 1907–1911, doi: 10.21437/Interspeech.2021-1694.

Rouas, J.-L., et Ioannidis, L. (**2016**). "Automatic Classification of Phonation Modes in Singing Voice : Towards Singing Style Characterisation and Application to Ethnomusicological Recordings," dans *Interspeech 2016*, pp. 150–154.

Rouas, J.-L., Shochi, T., Guerry, M., et Rilliard, A. (**2019**). "Categorisation of spoken social affects in Japanese : human vs. machine," dans *ICPhS*.

Schuller, B., Batliner, A., Bergler, C., Pokorny, F. B., Krajewski, J., Cychocz, M., Vollman, R., Roelen, S.-D., Schnieder, S., Bergelson, E., Cristia, A., Seidl, A., Warlaumont, A., Yankowitz, L., Nöth, E., Amiriparian, S., Hantke, S., et Schmitt, M. (**2019**). "The INTERSPEECH 2019 Computational Paralinguistics Challenge : Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," dans *Interspeech 2019*, doi: 10.21437/Interspeech.2019-1122.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., et Krajewski, J. (**2011**). "The INTERSPEECH 2011 Speaker State Challenge," dans *Interspeech 2011*, pp. 3201–3204, doi: 10.1.1.364.4935.

Shrout, P. E., et Fleiss, J. L. (**1979**). "Intraclass correlations : Uses in assessing rater reliability," Psychological Bulletin **86**(2), 420–428, doi: 10.1037/0033-2909.86.2.420.

Sjölander, K. (**2004**). "The Snack Sound Toolkit," Rapport Technique.

Smith, T., et Waterman, M. (**1981**). "Identification of common molecular subsequences," Journal of Molecular Biology **147**(1), 195–197, doi: 10.1016/0022-2836(81)90087-5.

Stolcke, A. (**2002**). "Srilm — An Extensible Language Modeling Toolkit," dans *Interspeech 2002*, pp. 901–904.

Vallat, R. (**2018**). "Pingouin : statistics in Python," Journal of Open Source Software **3**(31), 1026, doi: 10.21105/joss.01026.

Whitmore, J., et Fisher, S. (**1996**). "Speech during sustained operations," Speech Communication **20**(1-2), 55–70, doi: 10.1016/S0167-6393(96)00044-1.

Wu, H., Wang, W., et Li, M. (**2019**a). "The DKU-LENOVO Systems for the INTERSPEECH 2019 Computational Paralinguistic Challenge," dans *Interspeech 2019*, pp. 2433–2437, doi: 10.21437/Interspeech.2019-1386.

Wu, P., Rallabandi, S., Black, A. W., et Nyberg, E. (**2019**b). "Ordinal Triplet Loss : Investigating Sleepiness Detection from Speech," dans *Interspeech 2019*, pp. 2403–2407, doi: 10.21437/Interspeech.2019-2278.

Yeh, S.-L., Chao, G.-Y., Su, B.-H., Huang, Y.-L., Lin, M.-H., Tsai, Y.-C., Tai, Y.-W., Lu, Z.-C., Chen, C.-Y., Tai, T.-M., Tseng, C.-W., Lee, C.-K., et Lee, C.-C. (**2019**). "Using Attention Networks and Adversarial Augmentation for Styrian Dialect Continuous Sleepiness and Baby Sound Recognition," dans *Interspeech 2019*, pp. 2398–2402, doi: 10.21437/Interspeech.2019-2110.

Zhang, L., Duvvuri, R., Chandra, K. K. L., Nguyen, T., et Ghomi, R. H. (**2020**). "Automated voice biomarkers for depression symptoms using an online cross-sectional data collection initiative," Depression and Anxiety **37**(7), 657–669, doi: 10.1002/da.23020.

Cinquième partie

Classification automatique de la somnolence

Résumé

Dans cette partie, nous proposons trois façons différentes de traduire le problème clinique de la détection de la somnolence dans la voix en problème d'apprentissage automatique.

Dans le chapitre 16, nous présentons les concepts liés à la classification (validation croisée, métrique de performance ...) et la méthodologie employée pour concevoir et entraîner trois classifieurs sur trois symptômes liés à la somnolence, qui seraient prêts à être implémentés en utilisation réelle en suivant les principes des *Machine Learning Operations*.

Ensuite, dans le chapitre 17, nous confrontons la pertinence de classifier les symptômes et le raisonnement clinique, et nous proposons de classifier deux syndromes dérivés des trois symptômes classés précédemment.

Enfin, nous proposons dans le chapitre 18 un cadre théorique et méthodologique autour des réseaux de symptômes, qui ouvre des perspectives de recherche pour le futur développement de systèmes de diagnostic de pathologies à partir de la voix.

Mots-clés

Classification ; Symptômes ; Syndromes ; MLOps ; Réseaux de symptômes

Publications associées

Martin, V. P., Rouas, J.-L., Philip, P., Fourneret, P., Micoulaud-Franchi, J.-A., Gauld, C. (2022). How Does Comparison With Artificial Intelligence Shed Light on the Way Clinicians Reason ? A Cross-Talk Perspective. *Frontiers in Psychiatry*, 13.
https://doi.org/10.3389/fpsyt.2022.926286

Martin, V. P., Rouas, et Philip, P. (2022). Automatic detection of sleepiness-related syndromes through voice : are symptoms necessary ? [En cours d'évaluation par les pairs].

Chapitre 16

Élaboration d'un classifieur

Sommaire

16.1 Contexte et motivations

Le but des travaux présentés dans ce manuscrit est de proposer un réel outil utilisable par les cliniciens. Les efforts faits dans la partie précédente sur l'utilisation de marqueurs vocaux explicables aux médecins vont dans ce sens, mais ne sont pas suffisants. En effet, l'implémentation d'un système de classification pour son déploiement en conditions écologiques requiert un entraînement spécifique, qui est détaillé dans la section 16.2. Cette section introduit également les trois symptômes liés à la somnolence qui seront classifiés dans ce chapitre.

Nous présentons dans la section 16.3 les outils d'apprentissage automatique qui seront utilisés dans le système de classification, qui est décrit dans la section 16.4. Nous présentons les résultats de la procédure dans les sections 16.5 et 16.6.

Ensuite, nous faisons une étude approfondie des marqueurs sélectionnés par les classifieurs dans la section 16.7, et proposons quelques limites et perspectives de recherche sur ce sujet dans la section 16.8.

16.2 MLOps et tâches de classification

16.2.1 MLOps

L'objet de ce chapitre est de décrire l'élaboration de classifieurs de la somnolence, à l'aide uniquement des marqueurs vocaux précédemment décrits et de quelques données sur chaque patient. De la même façon que dans les chapitres précédents, nous nous restreignons ici à des classifications binaires (« Somnolent » ou « Non somnolent »).

Pour cela, nous utilisons une méthodologie robuste utilisée classiquement pour le déploiement de systèmes d'apprentissage automatique en situations réelles [*MLOps*, (Soh et Singh, 2020)]. Cette méthodologie est découpée en trois étapes :

1. une phase de double validation croisée, contenant deux boucles de validation croisée imbriquées l'une dans l'autre (*nested double cross-validation*), qui permet de sélectionner le système conduisant aux meilleures performances ;
2. une phase de simple validation croisée, qui permet de sélectionner les meilleurs hyperparamètres du système précédemment sélectionné ;
3. et enfin un apprentissage des paramètres, conduit sur l'ensemble de la base de données à notre disposition.

Une fois ces trois étapes réalisées, le système est prêt à être utilisé en conditions réelles. Nous en profitons alors pour analyser en profondeur les paramètres appris sur les descripteurs vocaux.

16.2.2 Tâches de classification

Dans cette partie, trois tâches de classification sont présentées :

— La classification d'une latence moyenne d'endormissement au TILE inférieure ou égale à 8 minutes. Cette tâche a pour but la mesure, à travers différents enregistrements audio d'un patient, d'estimer sa propension globale à l'endormissement diurne dans des conditions favorables au sommeil. Un classifieur ayant de bonnes performances permettrait, sous réserve d'autres études de validation, de remplacer les enregistrements EEG par des enregistrements de voix, qui pourraient à terme être délocalisés au domicile du patient.

— La classification d'un score à l'ESS supérieur à 15. Un classifieur performant sur cette tâche permet d'estimer la plainte des patients liée à leur tendance à l'endormissement dans la vie quotidienne.
— Enfin, la classification d'un score moyen des KSS remplies au cours d'un TILE supérieur à 5. Cette tâche permet d'estimer la moyenne de la somnolence estimée par les patients eux-mêmes, et permet ainsi de compléter les deux précédentes dimensions de la somnolence.

La figure 16.1 décrit le nombre de patients présentant chacun des trois symptômes introduits dans les trois définitions précédentes.

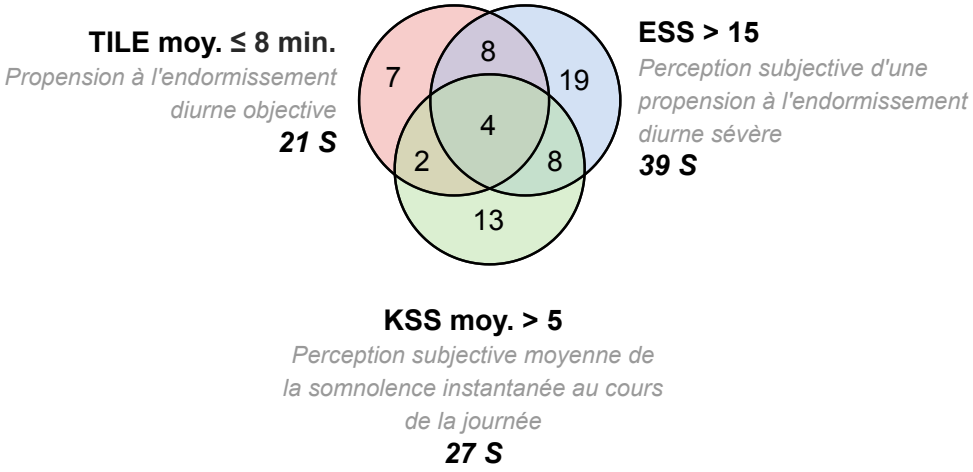

FIGURE 16.1 – Nombre de sujets présentant chacun des trois symptômes classifiés dans ce chapitre.

16.3 Validation croisée, hyperparamètres et paramètres

Ce chapitre introduit les différentes notions d'apprentissage automatique liées au système de classification proposé.

16.3.1 Validation croisée stratifiée à k blocs – *Stratified K-Fold*

Dans ce chapitre, nous utilisons une double validation croisée à blocs, illustrée dans la figure 16.2. Pour chaque itération externe, la base de données est divisée en N_{ext} sous-corpus. Un de ces sous-corpus est isolé pour former une base de test, et le processus est renouvelé pour séparer les échantillons restants en N_{int} sous-corpus, dont l'un est gardé comme sous-corpus de validation.

Ensuite, pour un jeu d'hyperparamètres de blocs donnés (cf. paragraphe suivant), les performances sont évaluées comme la moyenne des performances obtenues en faisant les N_{int} combinaisons possibles de corpus d'entraînement et de validation. Une fois le meilleur jeu d'hyperparamètres de bloc déterminé, les paramètres des blocs sont réentraînés et les performances de la boucle externe en cours sont calculées. La moyenne des performances calculées sur les différents sous-corpus de test permet d'estimer la performance moyenne d'un système pour un jeu d'hyperparamètres donnés du système.

Comparé à une validation croisée simple, le terme *stratifiée* signifie que la validation croisée utilisée est informée : les sous-corpus sont constitués de façon à avoir les mêmes ratios d'échantillons positifs et négatifs que la base de données entière, permettant un meilleur apprentissage des algorithmes.

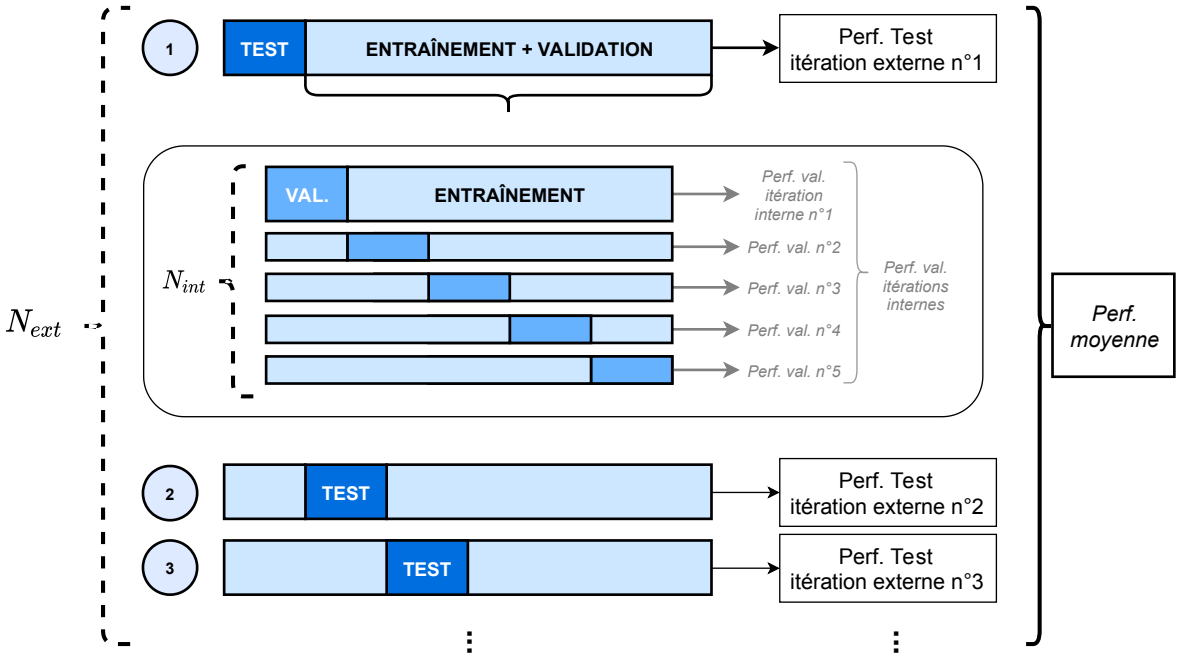

FIGURE 16.2 – Double validation croisée stratifiée à N_{ext} et N_{int} blocs (ici $N_{int} = 5$). *Perf.* : Performance, *val.* : validation.

Suivant comment sont distribuées les données dans la classe de somnolence, les paramètres optimaux de validation croisée peuvent différer. Afin de pouvoir comparer les systèmes entre eux, nous choisissons pour les trois tâches de classification la configuration la plus usuelle en MLOps ($N_{int} = 5$ et $N_{ext} = 10$) (Cearns *et coll.*, 2019; Kohavi et others, 1995).

16.3.2 Différents paramètres

L'utilisation d'une validation croisée double nécessite de clarifier les trois niveaux de paramètres du système qui seront estimés :

— Étape 1 : *Hyperparamètres du système* : ces paramètres servent à configurer les différents blocs composant le système de classification (la plupart du temps, est-ce que le bloc doit être utilisé ou non). Ces paramètres ne varient pas au cours de la double validation croisée : cette dernière a pour but d'estimer les meilleures performances atteignables avec un système figé. Ce sont ces paramètres qui sont estimés lors de la première phase du MLOps.

— Étape 2 : *Hyperparamètres de bloc* : chaque bloc du système est susceptible de posséder des hyperparamètres qui servent à configurer le bloc en lui-même (comme par exemple le type de classifieur utilisé, le seuil du bloc de sensibilité, les erreurs de STA filtrées ...). Ces hyperparamètres varient au cours de la boucle externe de validation croisée, et la boucle interne sert alors à estimer les meilleures performances qui peuvent être obtenues avec un jeu donné d'hyperparamètres de bloc, pour un système donné. Une fois les meilleurs *hyperparamètres de système* déterminés, ce sont ces hyperparamètres

qui sont estimés lors de la validation croisée simple (boucle interne) de l'étape 2 du MLOps.

— Étape 3 : *Paramètres* : certains blocs possèdent des paramètres internes (par ex. les poids de la régression logistique ou de l'analyse en composantes principales), qui sont évalués à chaque itération de la boucle interne de la validation croisée. Une fois les deux étapes précédentes achevées, ces paramètres sont figés lors de la troisième phase du MLOps, c'est-à-dire lors de l'entraînement du système final avec les hyperparamètres sélectionnés sur la base de données entière.

16.3.3 Évaluation des performances de classification

Toutes les tâches évaluées ici sont des classifications binaires, pour lesquelles les échantillons ont une valeur prédite et une valeur de vérité terrain qui vaut soit « Somnolent », soit « Non Somnolent ».

Matrice de confusion et taux de classification correcte

La façon la plus explicite de présenter des résultats de classification binaire est une matrice de confusion, dans laquelle sont indiqués sous forme de matrice 2×2 les Vrais Positifs (VP), Faux Positifs (FP), Vrais négatifs (VN) et Faux Négatifs (FN).

Cependant, si cette matrice ne laisse place à aucune ambiguïté sur les performances de classification d'un système, sa dimensionnalité ne permet pas d'en faire une métrique d'évaluation de performance pour comparer deux systèmes.

Dans la cadre d'une classification binaire, la fonction de performance la plus utilisée est la taux de classification correcte (*accuracy*). Celle-ci est définie par le ratio de la quantité de locuteurs correctement classifiés par le nombre total de locuteurs :

$$Acc = \frac{VP + VN}{VP + VN + FP + FN} \tag{16.1}$$

Si cette métrique est adaptée dans le cadre de problèmes équilibrés (c.-à-d. avec le même nombre de locuteurs étiquetés positifs et négatifs), elle échoue à rendre compte des performances d'un classifieur dans le cadre de problèmes déséquilibrés : dans ce dernier cas, un classifieur qui estimerait tous les échantillons comme appartenant à la classe majoritaire aurait malgré tout un haut score de classification.

> *Exemple :*
> Dans le cas de la classification du TILE considérée dans cette partie, un classifieur qui estimerait tous les locuteurs comme appartenant à la classe dominante (c.-à-d. la classe « Non somnolent ») atteindrait un score de classification correcte de 72/93 = 77.4%, sans avoir généralisé quelque concept que ce soit.

UAR et F1-score moyen biaisé

Il est donc nécessaire d'utiliser des fonctions de performances qui sont sensibles aux déséquilibres de classes. La métrique de référence utilisée lors des challenges IS11 et IS19 est le taux de rappel non pondéré – *Unweighted Average Recall* (UAR), défini comme la moyenne des taux de classification correcte sur chacune des classes de somnolence :

$$\text{UAR} = \frac{1}{2}\frac{VP}{VP + FN} + \frac{1}{2}\frac{VN}{VN + VP}$$

Avec cette métrique, un classifieur qui estime tous les échantillons comme appartenant à la même classe aurait ainsi un UAR de 50%.

Une autre métrique utilisée dans l'état de l'art que nous avons mentionnée dans les chapitres précédents est la moyenne pondérée des scores F1 (moyenne géométrique du rappel et de la précision) sur chaque classe :

$$\text{Rappel} = \frac{VP}{VP + FN}$$

$$\text{Précision} = \frac{VP}{VP + FP}$$

$$\begin{aligned} F1(P) &= 2\frac{\text{rappel (P)} \cdot \text{précision (P)}}{\text{rappel(P)} + \text{précision(P)}} \\ &= \frac{VP}{VP + 0.5(FP + FN)} \end{aligned}$$

Le F1-score moyen biaisé est alors défini par :

$$F1_{\text{moy}} = \frac{1}{N_P} F1(P) + \frac{1}{N_N} F1(N)$$

16.3.4 Moyenne des performances vs agrégation

Le nombre d'échantillons du corpus utilisé étant réduit, le nombre d'échantillons dans le sous-corpus de validation peut-être faible (17 échantillons avec $N_{ext} = 10$ et $N_{int} = 5$ sur une base de données de 93 locuteurs). En considérant en plus les contraintes d'équilibres de classes, les performances sont calculées sur un nombre trop réduit d'échantillons pour que cette valeur ait un réel sens une fois moyennée.

Nous préférons ainsi agréger, pour chaque itération de validation croisée, les classes estimées de chaque échantillon du sous-corpus de validation (boucle interne) ou du sous-corpus de test (boucle externe). La fonction de performance compare ensuite, une fois toutes les itérations de la boucle de validation ou de test effectuées, les valeurs ainsi agrégées et les vérités terrain correspondantes.

Exemple :
Prenons par exemple la boucle interne de la configuration $N_{ext} = 5$ et $N_{int} = 3$; et les vecteurs de prédiction ($\hat{y}$) et de vérité terrain (y) suivants (exemples fictifs) :

$$\hat{y}_1 = [0,1,1,1,1,0], y_1 = [0,1,1,1,1,1],\ UAR_1 = 75\%$$

$$\hat{y}_2 = [1,1,0,1,0,1], y_2 = [1,1,1,1,0,1],\ UAR_2 = 75\%$$

$$\hat{y}_3 = [1,0,0,0,0,1], y_3 = [1,1,0,1,0,1],\ UAR_3 = 75\%$$

En utilisant comme métrique la moyenne des UAR, nous obtiendrions alors une performance moyenne égale à $\mathbf{UAR_{moy} = 75\%}$.

En agrégeant les différentes prédictions d'une part et les vérités terrain d'autre part, nous avons :

$$\hat{y}_{agg} = [0,1,1,1,1,0,1,1,0,1,0,1,1,0,0,0,0,1]$$

$$y_{agg} = [0,1,1,1,1,1,1,1,1,1,0,1,1,1,0,1,0,1]$$

Nous obtenons alors : $\mathbf{UAR_{agg} = 63.75\%}$

Les arrondis successifs dans la calcul de la performance moyenne UAR_{moy} tendent ici à surestimer les performances réelles du système, correctement mesurées par la performance agrégée UAR_{agg}.

16.4 Conception du système de classification de la somnolence

Le système de classification utilisé dans ce chapitre est représenté dans la figure 16.3. Celui-ci est divisé en trois sous-parties :

— Un premier bloc faisant la fusion *a priori* des descripteurs vocaux extraits des enregistrements audio de la base de données ;
— Un deuxième bloc de décorrélation de ces descripteurs vis-à-vis de facteurs confondants comme l'âge, le sexe ou des comorbidités de la somnolence excessive ;
— Un troisième bloc de classification, sélectionnant les marqueurs les plus pertinents et classifiant le symptôme visé.

16.4.1 Entrée du système : fusion précoce des descripteurs

Descripteurs utilisés

Marqueurs acoustiques ($n = 74$) Dans ce chapitre, les marqueurs acoustiques utilisés sont calculés de deux manières différentes :

— les marqueurs présentés dans le chapitre 11 ($n = 44$), dont l'extraction est implémentée en Matlab, avec les boîtes à outils `Snack` pour l'extraction des segments voisés et de la fréquence fondamentale, et la boîte à outils `Covarep` pour l'estimation des formants ;
— les mêmes marqueurs, mais dont l'extraction est implémentée en Python et où tous les paramètres sont estimés avec la boîte à outils `Snack` ($n = 27$).

Erreurs des STA ($n = 112$) Le deuxième groupe de marqueurs utilisés comprend les erreurs des systèmes de transcription automatique de la parole (STA) décrits dans le chapitre 14.

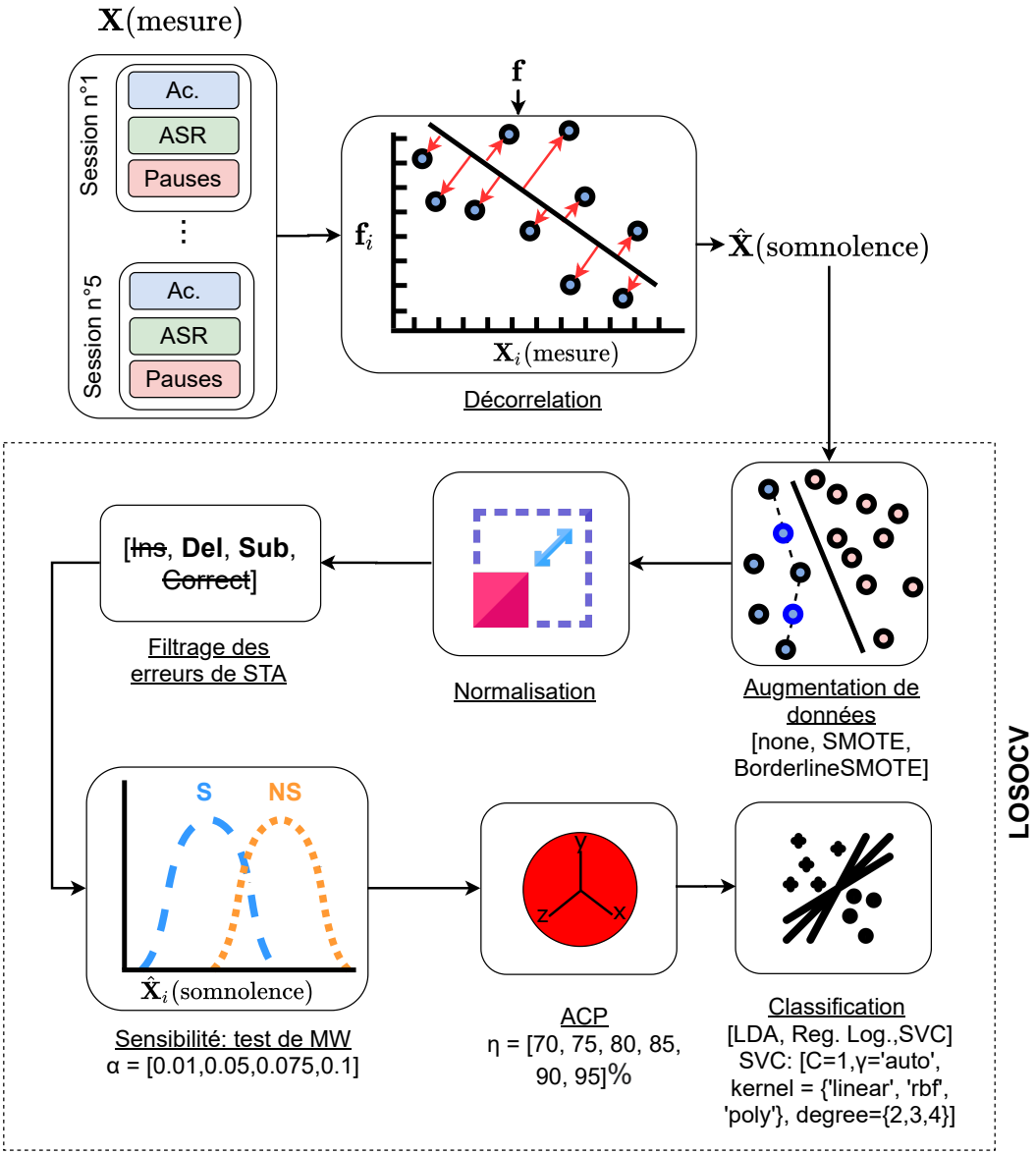

FIGURE 16.3 – Système de classification utilisé dans cette partie. STA : systèmes de transcription automatique ; *Ac.* : Acoustique, *ACP* : Analyse en composantes principales, LDA : *Linear Discriminant Analysis*, *Reg. Log.* : Régression Logistique, SVC : *Support Vector Classifier*.

Pauses ($n = 24$) Enfin, le dernier groupe de marqueurs mesure des caractéristiques du locuteur. Il comprend les statistiques des pauses de lecture présentées dans le chapitre 15 ($n = 14$), auxquelles on ajoute trois marqueurs calculés à partir du système de détection d'activité vocale (longueur moyenne, écart-type de la longueur et nombre de pauses) avant alignement avec l'hypothèse de transcription, quatre statistiques sur ces mêmes silences calculés avant alignement avec le détecteur d'activité vocale caractérisant les locuteurs sur leurs cinq enregistrements : la longueur médiane des pauses de lecture, l'écart-type et la longueur maximale des pauses.

On ajoute à ces marqueurs de pauses deux marqueurs caractérisant la vitesse de lecture : la longueur de chaque enregistrement, et la vitesse de lecture (en mots/minutes) du locuteur

sur chaque texte.

Fusion précoce des descripteurs

Notre objectif est d'estimer un trait du locuteur, qui est stationnaire sur des durées de plusieurs jours, semaines voire mois. À partir des cinq enregistrements d'un même sujet, trois stratégies peuvent être déployées pour estimer ces traits :

— une première version de *fusion tardive*, qui consisterait à faire un classifieur pour l'*état* des locuteurs (à l'échelle de la session de TILE), puis à appliquer une fonction de décision (par ex. le maximum ou la moyenne) sur les estimations associées aux cinq enregistrements de chaque locuteur. Cette méthode a l'avantage de bénéficier de beaucoup plus d'échantillons que de locuteurs à classifier, mais les résultats présentés dans le chapitre 12 semblent indiquer que le problème de classification des états est une tâche plus difficile que l'estimation des traits du locuteur ;

— une deuxième version de *fusion tardive*, qui consisterait à faire un classifieur par session du TILE, puis à appliquer une fonction de décision (introduite dans le chapitre 12). Cette version cumule cependant une tâche plus difficile (estimation des états), sans avoir les avantages de la démultiplication des échantillons dont bénéficie la première version de fusion tardive ;

— une *fusion précoce*, dans laquelle les descripteurs des différentes sessions sont agrégés pour ne former qu'un seul jeu de paramètres. Cette version permet de prendre en compte les effets de session (les mêmes descripteurs mesurés à deux sessions différentes sont considérés comme des paramètres différents), sans avoir à estimer les états du locuteur. C'est cette solution qui a été adoptée dans les chapitre 13 et 14, et que nous avons choisi d'implémenter dans ce chapitre.

Nous agrégeons donc les marqueurs des cinq siestes auxquels nous ajoutons la moyenne et l'écart-type de chaque marqueur au cours des siestes, ce qui conduit à 7 observations par descripteur et à une matrice de descripteurs en entrée du système de classification de dimension $(93, 1470)$.

16.4.2 Décorrélation

Le premier bloc du système est un processus de décorrélation entre les descripteurs audio et des co-facteurs qui peuvent interférer avec la voix et/ou la somnolence des patients, afin de ne conserver que les marqueurs spécifiques de la somnolence.

Les cofacteurs pris en compte sont les suivants :

— Sexe ;
— Âge ;
— IMC ;
— Tour de cou ;
— Niveau d'anxiété (HAD-A) ;
— Niveau de dépression (HAD-D) ;
— Niveau d'éducation (nombre d'années d'études après le brevet des collèges).

Pour chaque descripteur $\mathbf{X}_i(\text{mesure})$ $i \in \{1, ..., N_f\}$, nous notons $\mathbf{f}_i$ les cofacteurs qui corrèlent (ρ de Spearman, $p < 0.05$) avec $\mathbf{X}_i(\text{measure})$. Nous estimons la contribution de la somnolence dans la valeur de $\mathbf{X}_i$ par :

$$\hat{\mathbf{X}}_i(\text{somnolence}) = \mathbf{X}_i(\text{mesure}) - \hat{\mathbf{X}}_i(\mathbf{f}) \tag{16.2}$$

avec :

— $\hat{\mathbf{X}}_i(\text{somnolence})$ dénotant une estimation d'$\mathbf{X}_i(\text{somnolence})$ éventuellement influencée par des facteurs qui ne sont pas mesurés ;

— $\hat{\mathbf{X}}_i(\mathbf{f})$ dénotant une estimation de l'influence des facteurs externes sur $\mathbf{X}_i(\text{mesure})$.

Dans une première approche, nous estimons $\hat{\mathbf{X}}_i(f)$ grâce à une régression linéaire multivariée c'est-à-dire

$$\hat{\mathbf{X}}_i(\mathbf{f}) = \sum_k \alpha_k^i \mathbf{f}_k^i$$

Exemple : Prenons la fréquence fondamentale F0 sur la première session. Ce descripteur corrèle significativement (ρ de Spearman, $p < 0.05$) avec l'âge, le sexe, l'IMC, le tour de cou, le niveau d'éducation et l'HADD.
On estime donc $\hat{F}_0(\text{âge, sexe, IMC, niveau édu., HADD})$ par une régression linéaire multivariée, dont les coefficients sont donnés dans le tableau ci-dessous.
On estime ensuite $\hat{F}_0(\text{somnolence})$ par :

$$\hat{F}_0(\text{somnolence}) = F_0(\text{mesure}) - \hat{F}_0(\text{âge, sexe, IMC, niveau édu., HADD})$$

Le calcul des corrélations de $\hat{F}_0(\text{somnolence})$ avec les cofacteurs confirment la décorrélation de ce descripteur avec ces cofacteurs.

	Sexe	Age	IMC	Tour de cou	Niveau édu.	HADD	HADA
ϱ_{avant}	-0.29	-0.76	-0.34	-0.57	0.30	-0.27	0.14
(p)	(0.004)	(0.0)	(9.7e-4)	(1.7e-9)	(0.003)	(0.009)	(0.17)
α_{reg}	-0.57	-75.5	-0.12	-0.13	0.19	-1.22	-
$\varrho_{\text{après}}$	0.06	-0.12	0.02	-0.07	0.06	0.03	0.15
(p)	(0.58)	(0.26)	(0.79)	(0.46)	(0.54)	(0.32)	(0.14)

Après cette étape de décorrélation, nous éliminons tous les marqueurs corrélant encore avec un des co-facteurs contrôlés dans cette étude.

Cette étape est appliquée de manière identique à tous les systèmes évalués, et la matrice de descripteurs en entrée des trois systèmes est identique.

16.4.3 Description des différents blocs du système de classification

Augmentation des données

Les classes sont déséquilibrées sur les trois problèmes considérés : 22.5% de positifs pour la détection du TILE, 42% de positifs pour la détection de l'ESS et moins de 37% de positifs pour la détection de la KSS moyenne. Ces déséquilibres peuvent être compensés de deux façons :

— En souséchantillonnant la classe dominante, c'est-à-dire en éliminant des échantillons de la classe la plus représentée ;

— Ou en suréchantillonnant la classe minoritaire, c'est-à-dire en créant de nouveaux échantillons ressemblant à ceux de la classe minoritaire.

La base de données utilisée ayant déjà une taille faible, nous choisissons de suréchantillonner la classe minoritaire avec des algorithmes dérivés de SMOTE [*Synthetic Minority Over-Sampling TEchnique*, (Chawla *et coll.*, 2002)].

Nous considérons deux algorithmes ici, implémentés dans la boîte à outils Python `imblearn` :

— l'algorithme *SMOTE*, qui génère des échantillons se trouvant aléatoirement sur de segments reliant deux échantillons réels de la classe minoritaire ;

— l'algorithme *BorderlineSMOTE*, qui procède de même en se concentrant sur les limites entre classes dans chaque dimension, afin d'augmenter le contraste dans les zones de limite interclasses.

L'algorithme utilisé (ou sa non utilisation) est un hyperparamètre du système, qui est fixe pour les deux boucles de validation croisée.

Normalisation

La deuxième étape du système de classification est la normalisation des descripteurs (*z-normalization*, centrage et réduction à une moyenne $\mu = 0$ et un écart-type $\sigma = 1$).

Filtrage des erreurs des STA

Nous avons pu observer dans le chapitre 14 que les insertions et substitutions produites par les STA étaient plus adaptées à la détection de la latence d'endormissement pathologique au TILE tandis que les substitutions et délétions étaient plus adaptées à la détection de la plainte sévère de SDE.

Ce bloc permet ainsi de tester le système avec toutes les combinaisons possibles d'erreurs de STA, afin de sélectionner celle qui permet le meilleur score de classification. La présence – ou non – de ce bloc est un hyperparamètre du système (et il est donc fixe pour les deux boucles de validation croisée), tandis que les différentes combinaisons d'erreurs testées sont un hyperparamètre du bloc de filtrage des erreurs de STA, qui change donc au sein de la boucle externe de validation croisée.

Pouvoir discriminant des descripteurs

Ce bloc permet de sélectionner les marqueurs en fonction de leur sensibilité à la somnolence, opérationnalisée par un test de Mann-Whitney bilatéral entre les descripteurs des deux classes S et NS sur la base d'entraînement. Les marqueurs pour lesquels la *p-value* du test est supérieure à un seuil α sont exclus, jugés insuffisamment sensibles à la somnolence.

La présence – ou non – de ce bloc est un hyperparamètre du système (et il est donc fixe pour les deux boucles de validation croisée), tandis que le seuil α est choisi parmi les valeurs suivantes à chaque itération de la boucle externe de validation croisée : $\alpha \in \{0.01, 0.05, 0.075, 0.1\}$.

Analyse en composantes principales

Certains classifieurs (comme par exemple les régressions logistiques) nécessitent que les données qu'elles traitent aient des dimensions qui soient orthogonales entre-elles. À cette fin, un des blocs du système est une analyse en composantes principales, dont le but est à la fois de réduire la dimensionnalité des données, mais aussi d'orthogonaliser les descripteurs vocaux précédemment sélectionnés entre eux.

La présence – ou non – de ce bloc est un hyperparamètre du système, et il est donc fixe pour les deux boucles de validation croisée, tandis que le ratio de variance conservé η est un hyperparamètre du bloc choisi parmi les valeurs suivantes : $\eta \in \{70\%, 75\%, 80\%, 90\%, 95\%\}$ Les paramètres de l'ACP (vecteurs propres et coefficients de projection) sont estimés à chaque itération de la boucle interne de la validation croisée.

Classification

Enfin, le dernier bloc du système est un classifieur, choisi parmi un des trois classifieurs suivants :

— *Une régression logistique* : ce classifieur, le plus élémentaire possible, est l'équivalent d'une régression linéaire sur les étiquettes binarisées, à laquelle une fonction *sigmoïde* est ensuite appliquée. Ayant fait ses preuves dans la classification de la somnolence grâce aux erreurs des STA (cf. chapitre 14), ce classifieur est un candidat plausible pour la classification de la somnolence à partir de tous les descripteurs conçus. Ce classifieur n'a pas d'hyperparamètre.

— *Un Séparateur à Vastes Marges* (*Support Vector Machine* – SVM) : classifieur le plus utilisé dans l'état de l'art de la classification de la somnolence instantanée et ayant fait ses preuves dans la classification de la somnolence au long court à partir des descripteurs acoustiques et des erreurs de lecture (cf. chapitre 13), ce classifieur est l'équivalent d'une régression linéaire pour laquelle l'écart interclasses sur chaque dimension est maximisé (d'où l'appellation de « vastes marges »). Ce classifieur a entre deux et quatre hyperparamètres, qui sont le type de noyau utilisé (linéaire, gaussien ou polynomial), le facteur de correction C (fixé à la valeur par défaut de 1 dans notre système), le facteur de dilatation γ (pour les noyaux gaussiens et polynomiaux uniquement, fixé à la valeur par défaut $1/N_{\text{descripteurs}}$, noté « auto » dans la suite), et enfin le degré du polynôme (uniquement dans le cas d'un noyau polynomial, choisi parmi les valeurs 2, 3 ou 4).

— *Une analyse discriminante linéaire* (*Linear Discriminant Analysis* – LDA) : classifieur qui estime des distributions conditionnelles gaussiennes pour chacune des classes et procède ensuite à la classification. Ce classifieur n'a pas d'hyperparamètre.

Ces trois classifieurs sont configurés de façon à biaiser l'importance des échantillons d'entraînement par l'effectif de la classe à laquelle il appartient. Ainsi, un échantillon appartenant à une classe largement dominante influera moins la mise à jour des poids qu'un échantillon appartenant à la classe dominée.

Le classifieur utilisé est un hyperparamètre du système, et ne varie donc pas au cours de la double validation croisée. Les hyperparamètres du SVM sont des hyperparamètres de bloc, et varient donc au cours de la validation croisée externe. Enfin, ces trois classifieurs possèdent des paramètres internes (poids des régressions ou des projections) qui sont estimés à chaque itération de la boucle interne de la double validation croisée.

16.5 Résultats – Étape n°1 : Sélection du modèle

Les hyperparamètres des systèmes produisant les meilleures performances pour chaque tâche sont rapportés dans le tableau 16.1.

Les performances ne sont reportées dans ce tableau qu'à titre indicatif : seule leur valeur relative aux autres systèmes est importante pour la sélection des meilleurs systèmes. Ainsi, même si le système de détection de la plainte de SDE sévère (ESS) ne dépasse pas les 60%, la configuration maximisant l'UAR sur la double validation croisée est gardée pour la détermination des meilleurs hyperparamètres de blocs, étape pour laquelle les performances sur la validation croisée simple sont habituellement reportées dans la littérature.

Ref.	Tâche	SMOTE	Filtre ASR	Sen.	ACP	Class.	UAR Test (%)
I	TILE ≤ 8.0	-	✓	×	✓	Rég. Log.	64.38%
II	ESS > 15	BorderlineSMOTE	✓	×	×	LDA	58.48%
III	KSS > 5	BorderlineSMOTE	×	×	✓	Rég. Log.	65.4%

Tableau 16.1 – Hyperparamètres des systèmes donnant les meilleures performances sur une double validation croisée.

Ref.	Err. ASR	MW α	PCA η	Classif	UAR CV	UAR oracle
(I)	[Del, Sub]	-	75%	Rég. Log.	57.9%	85.5%
(II)	[Correct]	-	-	LDA	62.4%	85.3%
(III)	-	-	95%	Rég. Log.	61.8%	100%

Tableau 16.2 – Hyperparamètres de blocs donnant les meilleures performances sur les systèmes précédemment sélectionnés (validation croisée simple).

16.6 Résultats – Étape n°2 : Sélection des meilleurs hyperparamètres de bloc

Une fois les systèmes sélectionnés, la deuxième étape est la sélection des meilleurs hyperparamètres de bloc. Pour cela, nous procédons à une validation croisée simple, en testant toutes les combinaisons d'hyperparamètres de bloc possibles. Les résultats de cette recherche sont rapportés dans le tableau 16.2.

Les performances obtenues sont plus faibles que celles obtenues jusqu'à présent, sur les systèmes n'utilisant qu'un seul type de marqueurs (cf. chapitres 12 et 14). Ces faibles performances sont discutées dans la section 16.8.

16.7 Résultats – Étape n°3 : Interprétation des marqueurs vocaux sélectionnés

La dernière étape du MLOps consiste à entraîner le système sélectionné avec les meilleurs hyperparamètres de bloc précédemment estimés sur l'intégralité de la base de données utilisée. Une fois cette étape achevée, le système est prêt à être déployé en conditions réelles. Les performances du système entraîné et testé sur l'intégralité de la base de données (oracle) sont rapportées dans la dernière colonne du tableau 16.2.

Dans cette section, nous profitons de l'interprétabilité des blocs inclus dans les systèmes sélectionnés pour investiguer les marqueurs sélectionnés et leurs importances dans la classification.

De même que pour le chapitre 14, nous proposons d'étudier dans cette section les poids

attribués par les classifieurs et/ou l'ACP aux différents descripteurs ayant passé les étapes de sélection des marqueurs. Pour cela, nous proposons deux analyses.

16.7.1 Analyse des marqueurs contribuant le plus à la classification

Méthode

Dans la figure 16.4, nous proposons des analyses similaires à celles proposées dans le chapitre 14 : pour les systèmes classifiant la latence d'endormissement moyenne au TILE et la somnolence moyenne (KSS moy.), nous représentons les cinq composantes de l'ACP ayant les poids les plus importants (en valeur absolue) dans la régression logistique, et une synthèse des dix marqueurs ayant les poids les plus importants (en valeur absolue) dans ces composantes. Pour la classification de la SDE sévère (ESS), le système n'incluant pas de bloc d'ACP, nous proposons une analyse similaire directement sur les coefficients de la LDA.

Résultats

Pour les trois systèmes, les dimensions les plus influentes sur la classification sont des erreurs des STA.

Système I : TILE $\leq$ 8 Pour le système I, les erreurs sélectionnées sont les délétions et les substitutions, tandis que dans le chapitre 14 précédent, les erreurs reliées à la somnolence objective étaient respectivement les substitutions dans l'étude statistique de la section 14.3 et les substitutions et insertions dans la section 14.5 : ces différentes observations tendent toutes vers un lien étroit entre substitutions et latence moyenne au TILE.

Parmi les marqueurs utilisés par ce système, les scores et nombres de pauses négatives sur la troisième sieste sont parmi les marqueurs les plus importants (4e dimension la plus importante). Ces observations sont cohérentes avec les résultats de l'analyse de la section 15.4, qui avait révélé une influence des combinaisons TILE : IMC : ESS et TILE : ESS : Socio-Dem. avec respectivement N^- et S^-.

Enfin, deux marqueurs acoustiques isolés (l'énergie moyenne sur la deuxième session et l'écart-type sur les sessions de la fréquence du 4e formant) sont inclus dans les 4e et 5e dimensions les plus importantes, mais ne permettent pas de définir de tendance sur ce type de marqueurs. Une étude plus fine est requise pour déterminer s'il s'agit d'une contribution réelle de ces marqueurs ou d'artefacts statistiques.

Système II : ESS $>$ 15 Pour le système II, en plus des erreurs des STA (uniquement des statistiques liées au nombre d'unités correctement décodées), les marqueurs les plus importants sont des marqueurs acoustiques, et notamment des statistiques liées aux formants (amplitude et bande passante) et le ratio de parties voisées, qui sont cohérentes avec les résultats observés précédemment dans la section 14.4. Ces marqueurs sont certainement plus présents que dans les autres systèmes en raison du faible nombre d'erreurs de STA filtrées (uniquement les unités correctement décodées).

Concernant les erreurs des STA, la section 14.3 avait lié SDE et insertions, tandis que la section 14.4 classifiait la SDE sévère en se basant sur des délétions et substitutions. Le système II ne se base sur aucune des combinaisons précédemment observées : une étude plus approfondie de ces marqueurs est nécessaire.

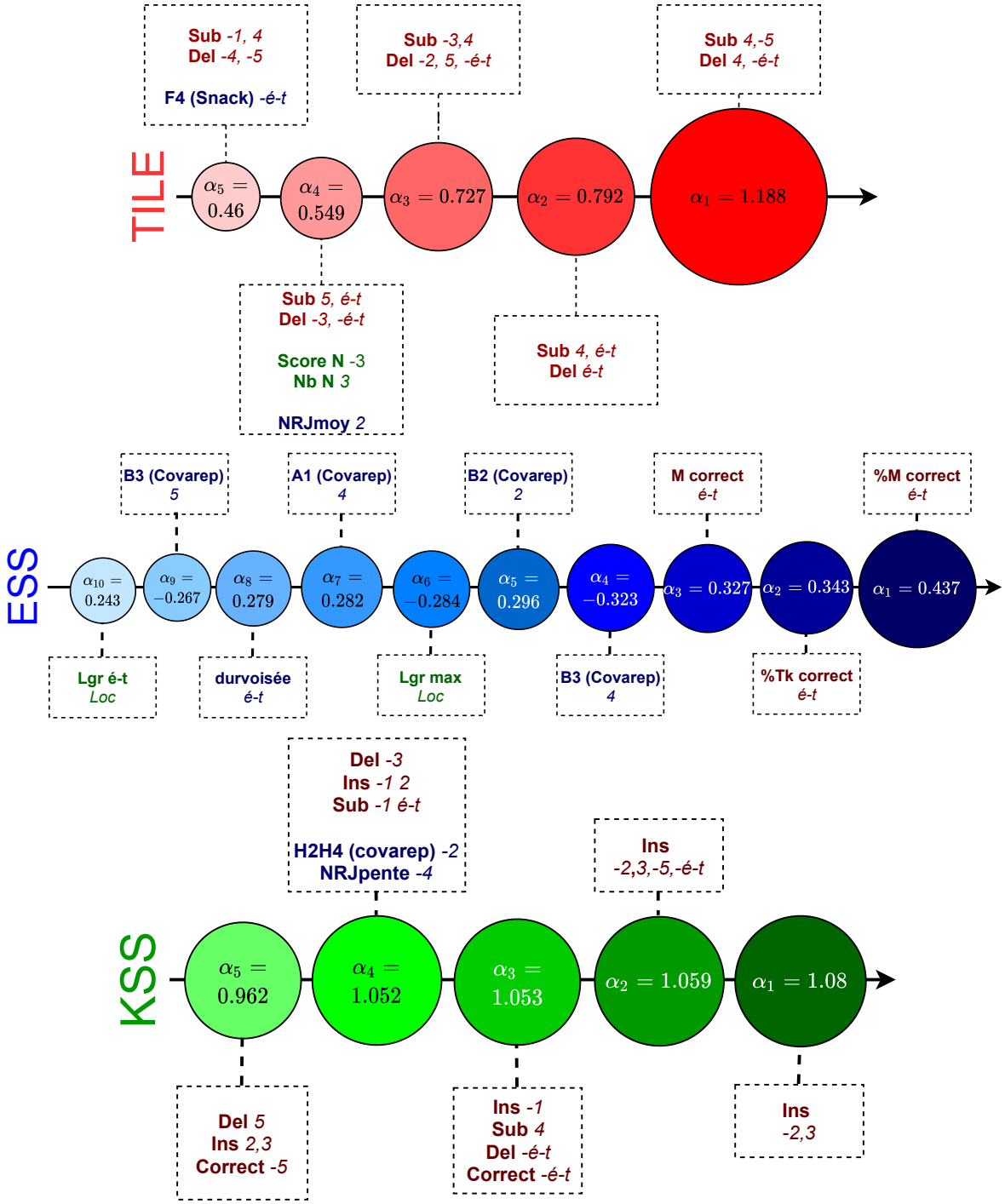

FIGURE 16.4 – Marqueurs ayant les plus de poids dans la classification des trois symptômes étudiés. Les signes négatifs correspondent à une contribution négative dans l'ACP.

Enfin, deux marqueurs liés aux pauses de lecture au niveau global du locuteur (écart-type et longueur maximale des pauses) font partie des 10 marqueurs les plus importants pour ce classifieur.

Système III : KSS $>$ 5 Enfin, pour le système III, les marqueurs dominants sont tous liés à des erreurs des STA.

Il est à noter que les deux dimensions les plus importantes sont liées aux insertions faites par les STA, ce qui est en phase avec le résultat de la section 14.3, mais que les 3 autres dimensions ayant le plus de poids dans la classification sont dirigées par un mélange de divers type d'erreurs des STA.

De même que pour le système I, deux marqueurs acoustiques isolés (H2H4 sur la deuxième sieste et la pente de l'énergie sur la 4e sieste) sont inclus dans une des cinq dimensions prédominantes, et nécessiteraient une étude plus complète afin de dégager une réelle tendance de la modification acoustique de la voix des sujets étant somnolents au cours de la journée.

16.7.2 Analyse par session et par catégorie de marqueurs

Méthode

Une deuxième analyse que nous proposons ici concerne les importances relatives des différentes catégories de marqueurs et des différentes sessions. Pour chaque combinaison de session et de catégorie, nous additionnons les valeurs absolues de tous les poids dans le système II et les valeurs absolues des poids du classifieur pondérés par les poids de l'ACP pour les systèmes I et III. Ces valeurs sont ensuite exprimées en ratio du poids total obtenu, permettant d'identifier quels sont les domaines ayant le plus de poids dans la décision du classifieur.

Résultats

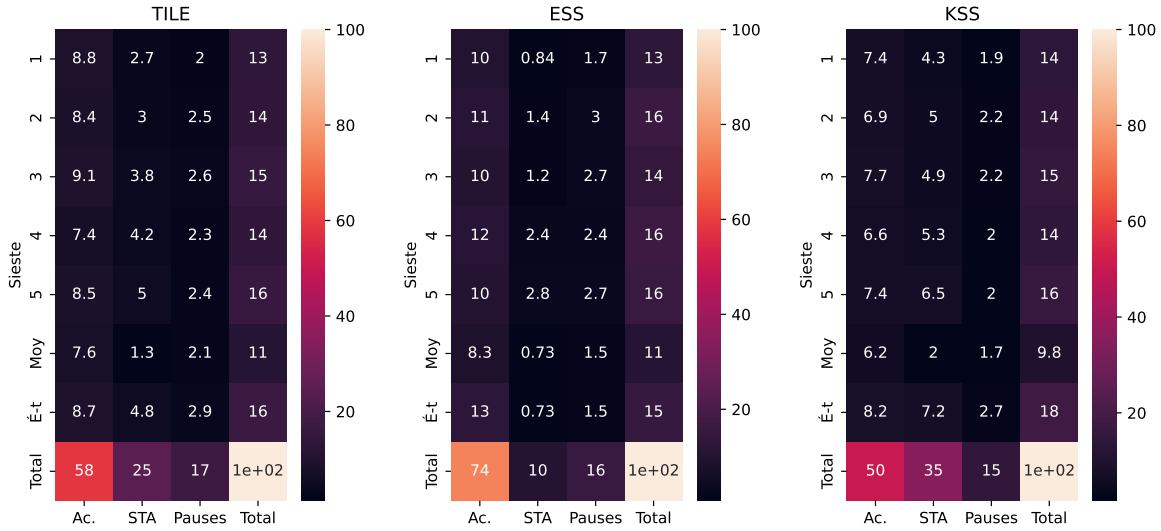

FIGURE 16.5 – Distribution relative du poids total absolu attribué aux marqueurs par les classifieurs en fonctions de l'itération du TILE et du groupe de marqueurs vocaux.

Nous rapportons ainsi dans la figure 16.5 les pourcentages de poids absolu total en fonction des catégories de marqueurs et des sessions.

Les marqueurs acoustiques sont dominants pour les trois systèmes, avec une dominance plus marquée pour le système II (41% du poids absolu total), que nous supposons être une conséquence du filtre des erreurs de STA qui est plus sélectif pour ce système, et qui laisse donc plus d'importance aux autres types de marqueurs. Pour les systèmes I et III, les erreurs

des STA sont les deuxièmes marqueurs les plus importants (avec respectivement 25 et 35% des poids absolus), et les pauses les troisièmes (avec respectivement 17 et 15% de poids absolu total), tandis que pour le système II, les pauses sont le deuxième marqueur le plus important (16% de poids absolu total) et les erreurs des STA le troisième (10% du poids absolu total).

Concernant l'importance des sessions, la moyenne des marqueurs au cours des sessions est la moins représentée dans les trois systèmes, alors que l'écart-type est une modalité fortement représentée dans les trois systèmes : les variations des marqueurs au cours des siestes semblent mieux refléter la somnolence au long cours que leur moyenne. Cependant, les différences de contribution entre les sessions sont faibles et ne permettent pas de dégager une tendance franche sur cette question.

16.8 Limites et perspectives

16.8.1 Paramètres de validation croisée N_{int} et N_{ext}

Les paramètres de validation croisée N_{int} et N_{ext} sont très discutés dans la littérature (Wong, 2015; Marcot et Hanea, 2021; Kohavi et others, 1995; Vanwinckelen et Blockeel, 2012; Fushiki, 2011; Bengio et Grandvalet, 2003; Anguita *et coll.*, 2012). Les valeurs utilisées habituellement fluctuent entre 3 et 10 – avec un engouement particulier pour cette dernière valeur, explicitement recommandée dans l'article de référence (Kohavi et others, 1995). Cette dernière étude menée sur sept bases de données précise pourtant dans sa conclusion :

> « Nos résultats indiquent que pour des bases de données similaires aux nôtres, la meilleure méthode à utiliser pour la sélection des modèles est une validation croisée stratifiée à 10 blocs » [1]

Ainsi, il ne semble pas y avoir de règle générale sur le choix de ces paramètres, celui-ci étant conditionné aux données sur lesquelles l'entraînement est fait.

Les trois problèmes étudiés dans ce chapitre ayant des géométries différentes, une recherche élargie en incluant les paramètres N_{ext} et N_{int} comme hyperparamètres de systèmes pourrait conduire à des résultats plus prometteurs que ceux obtenus avec les paramètres utilisés dans ce chapitre.

16.8.2 Décorrélation des descripteurs

Le bloc de décorrélation proposé dans la section 16.4 n'est qu'une première approche à un problème beaucoup plus vaste. En effet, nous avons fait le choix de décorréler les marqueurs avec les cofacteurs présents dans la base TILE sur l'ensemble de la base données, entraînant un possible surapprentissage des relations entre les descripteurs et ces cofacteurs. Cependant, cette implémentation est un premier effort pour prendre en compte la spécificité des descripteurs vis-à-vis d'un ensemble de cofacteurs, ce qui nécessiterait une base de données beaucoup plus importante pour permettre un apprentissage rigoureux.

Par ailleurs, si les relations entre descripteurs et cofacteurs sont sujettes à un surapprentissage, la méthodologie suivie pour la détection de la somnolence a été menée de manière rigoureuse (double validation croisée) : la connaissance apportée par cette étape de décorrélation est indépendante du problème central traité ici, de détection de la somnolence dans la voix.

1. "Our results indicate that for real-word datasets similar to ours, the best method to use for model selection is ten-fold stratified cross validation even if computation power allows using more folds"

Enfin, il est à noter que l'estimation de cette relation (régression linéaire multivariée) n'est qu'une première approche simpliste, et nécessiterait des investigations plus poussées. Une approche récente trouvée dans la littérature pour différencier les caractéristiques propres du locuteur de son état pathologique est l'utilisation, dans le cadre de systèmes utilisant l'apprentissage profond, d'une fonction de coût de comparaison (la *triplet loss*), utilisée avec succès par exemple pour la détection des troubles bipolaires par la voix (Du *et coll.*, 2018) ou même la somnolence (Wu *et coll.*, 2019).

16.8.3 Et l'*apprentissage profond* ?

Enfin, il est à noter qu'aucun des systèmes proposés ni dans ce chapitre ni dans les précédents (ni suivants) n'utilise de technique basée sur l'apprentissage profond. En effet, ce choix de notre part s'appuie sur deux arguments :

D'une part, la supériorité de l'apprentissage profond sur les techniques d'apprentissage automatique « classique » n'est une réalité que pour certaines tâches particulières : dans une majorité de tâches liées à la médecine et à la santé, les techniques d'apprentissage « classiques » offrent des performances comparables aux réseaux de neurones profonds (Christodoulou *et coll.*, 2019; Desai *et coll.*, 2020; Lynam *et coll.*, 2020; Nusinovici *et coll.*, 2020; Gravesteijn *et coll.*, 2020; Cho *et coll.*, 2021), en apportant finalement que peu de nouveautés (Faes *et coll.*, 2022). C'est d'ailleurs une des conclusions du challenge IS19, pour lequel le vainqueur, en compétition avec des systèmes utilisant des systèmes d'apprentissage profond parfois imposants, n'utilise que des descripteurs explicites (*hand-crafted features*) avec des SVM (cf. chapitre 11).

D'autre part, nous nous sommes limités à des techniques d'apprentissage « classiques » en raison de leur bonne interprétabilité : alors qu'un sujet émergent de recherche – *l'explainable AI* – concentre les efforts de nombreuses équipes de recherche de par le monde, nous embrassons la doctrine de (Rudin, 2019) qui prône l'utilisation de modèles intrinsèquement explicables, plutôt que l'investissement d'efforts colossaux dans l'interprétation *a posteriori* de modèles très complexes. Cette interprétabilité est par ailleurs un point critique de notre travail : la collaboration avec les médecins du sommeil ne tolère pas, comme c'est le cas dans une partie des systèmes de l'état de l'art, que les descripteurs soient extraits de manière non explicite, sans garantie d'interprétabilité.

16.9 Conclusion

En conclusion, nous avons appliqué les préceptes du *MLOps* pour entraîner des classifieurs avec l'objectif de les rendre disponibles pour une utilisation en conditions réelles. Cependant, la procédure nécessite une double boucle de validation croisée, pour laquelle nous pensons que nos données ne sont pas assez nombreuses pour évaluer correctement les différents hyperparamètres de système et de bloc, ce qui est responsable des faibles performances de classification obtenues. Nous avons malgré tout proposé une interprétation des marqueurs sélectionnés par ces systèmes, qui semblent cohérents avec les résultats obtenus dans les précédents chapitres.

Dans la partie suivante, nous proposons d'étudier une nouvelle tâche de classification, non pas basée sur des *symptômes* liés à la somnolence, mais sur des *syndromes* construits à partir des ceux-ci.

Chapitre 17

Du *symptôme* au *syndrome*

Sommaire

17.1 Contexte et objectif

À la fois dans les compétitions internationales IS11 et IS19 (précisément décrites dans le chapitre 11) et dans les systèmes de classification proposés dans les chapitres précédents, le but est d'estimer ou de classifier *un score de somnolence*, mesuré soit par le score à un questionnaire de somnolence (ESS, KSS), soit par le résultat à un test objectif (TILE). Cependant, dans leur pratique clinique courante, les cliniciens utilisent rarement ces mesures de façon aussi stricte que dans la façon dont les problèmes d'apprentissage automatiques sont formulés (Blashfield et Herkov, 1996; Zimmerman et McGlinchey, 2008; Bowen, 2006; Bostic et Rho, 2006). En effet, comme le précise Norman dans (Norman, 2000) :

> « [. . .] le clinicien essayant de faire un diagnostic réfléchit quasiment exclusivement à l'échelle du syndrome » [1]

Il définit un *syndrome* comme « un ensemble de signes et symptômes » [2] de plus haut niveau conceptuel que ceux-ci, ayant une forme de cohérence pour le clinicien.

Dans ce chapitre, nous avons pour objectif de changer la formulation du problème de classification de la somnolence tel qu'il est posé habituellement pour le rapprocher du raisonnement clinique. Pour cela, nous proposons d'estimer deux *syndromes* liés à la somnolence au long cours – la perception subjective de somnolence excessive et la propension sévère à l'endormissement diurne – dérivés de trois symptômes : la mesure objective de la propension à l'endormissement diurne (latence moyenne au TILE), la perception subjective de somnolence diurne excessive sévère (ESS), et l'évaluation subjective moyenne de la somnolence au cours de la journée (KSS moyenne). Ce faisant, nous rapprochons la formulation du problème d'apprentissage automatique du raisonnement clinique, facilitant son adoption par les cliniciens, alors familier de la tâche effectuée par le système.

Pour cela, nous procédons de deux manières distinctes.

— Méthode n°1 : nous entraînons des systèmes de classification dédiés à partir des marqueurs vocaux, sur des données annotées avec le statut de chaque locuteur concernant le *syndrome* étudié ;
— Méthode n°2 : nous estimons dans un premier temps les *symptômes* avec les systèmes présentés dans le chapitre précédent, et procédons ensuite à la fusion *a posteriori* des estimations obtenues pour détecter les *syndromes*.

Nous faisons l'hypothèse que la méthode n°1 fonctionnera mieux, du fait qu'elle ne repose pas sur une première estimation correcte des symptômes, et évite ainsi des approximations en cascade. Cependant, estimer les symptômes, ce qui est nécessaire pour la méthode n°2, permettrait d'estimer d'autres syndromes basés sur ces symptômes, voire de proposer une autre approche de la pathologie basée sur ceux-ci (cf. chapitre 18 suivant).

Ce chapitre est structuré comme suit. Dans la section 17.2, nous présentons plus en détail les deux syndromes que nous désirons classifier, ainsi que les trois symptômes de somnolence à partir desquels ils sont définis. Les systèmes de classification et la méthodologie de validation croisée sont présentés dans la section 17.3. Les résultats des deux méthodes de classification de syndrome sont présents dans la section17.4, que nous discutons dans la section 17.5. Enfin, nous proposons quelques conclusions dans la section 17.6.

1. "the clinician attempting to make a diagnosis is dealing almost exclusively at the syndrome level"
2. "a cluster of signs and symptoms"

17.2 *Symptômes* vs *syndromes*

17.2.1 Corpus

Le corpus utilisé dans ce chapitre est le corpus TILE-93, présenté dans le chapitre 7.

17.2.2 Définition des symptômes

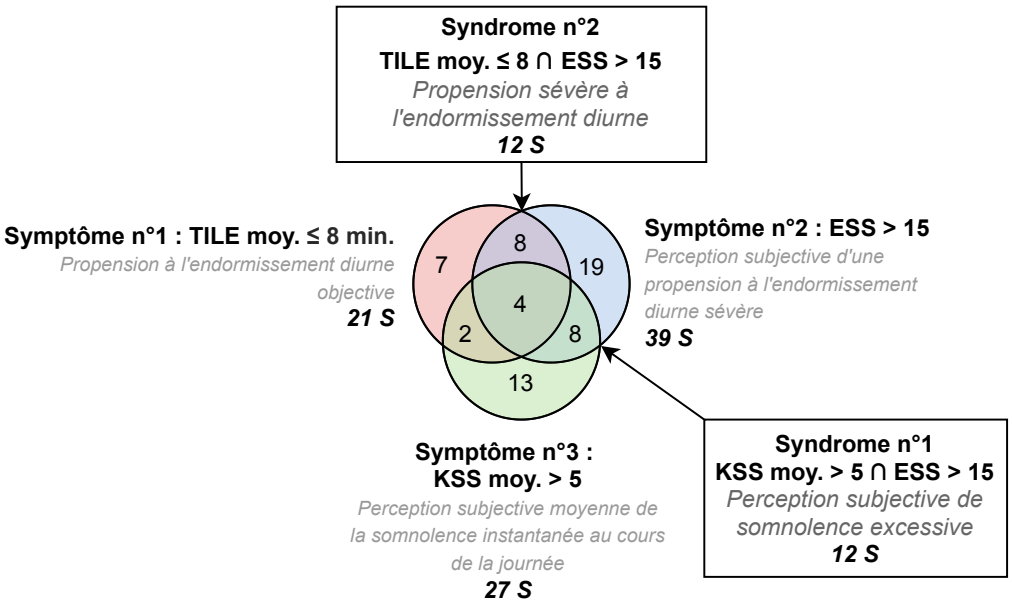

FIGURE 17.1 – Nombre de patients atteints des symptômes et syndromes classifiés dans ce chapitre.

Les symptômes et syndromes que nous désirons classifier sont représentés dans la figure 17.1. Nous nous limitons, comme dans le chapitre 16 précédent, à trois symptômes liés à la somnolence au long cours dont les mesures sont incluses dans la base TILE.

Symptôme n°1 : Mesure objective de la propension à l'endormissement diurne dans des conditions favorables au sommeil (TILE $\leq$ 8)

La propension à l'endormissement diurne dans des conditions favorables au sommeil est mesurée par la latence moyenne d'endormissement au TILE. Une latence moyenne inférieure ou égale à 8 est habituellement considérée comme pathologique (Arand *et coll.*, 2005) : comme dans le reste de ce document, nous choisissons cette valeur pour séparer les deux classes de notre tâche de classification binaire. Cela conduit à un problème de classification binaire déséquilibré, avec seulement 21 patients présentant ce symptôme (22.6% du corpus).

Symptôme n°2 : Perception subjective de somnolence diurne excessive sévère

La perception subjective de somnolence diurne excessive est habituellement mesurée par un score supérieur à 10 à l'ESS. Cependant, comme dans les chapitres 14 et 16 précédents, cette limite ne convient pas à l'apprentissage correct d'un classifieur : plus de 87% du corpus est au-dessus de cette limite puisque notre population est composée de patients atteints de diverses formes d'hypersomnies. En conséquence, nous considérons dans ce chapitre une limite de 15,

utilisée précédemment comme limite pour différencier la SDE *sévère* (Philip *et coll.*, 2008). Cela conduit à un problème de classification binaire quasiment équilibré avec 39 patients présentant une plainte de SDE *sévère* (41.9% du corpus).

Symptôme n°3 : Évaluation subjective moyenne de la somnolence au cours de la journée

L'évaluation subjective moyenne de la somnolence au cours de la journée est mesurée dans cette partie comme la moyenne des cinq KSS proposées aux patients durant le passage de leur TILE (un avant chaque sieste). Durant la compétition IS11, une valeur limite de 7.5 était utilisée pour la détection de la somnolence à court terme, ce qui était en adéquation avec l'association entre KSS supérieures à 7 et la baisse de performances observée (Kaida *et coll.*, 2006).

Cependant, puisque la tâche considérée ici est l'estimation de la somnolence moyenne, nous choisissons un seuil de 5, correspondant au niveau neutre de la KSS : les niveaux au-dessus de 5 ont tous le mot « somnolent » dans leur description, alors qu'un score inférieur à 5 correspondant à un patient éveillé. Pour cette tâche de classification binaire, 27 patients (29% du corpus) présentent ce symptôme.

17.2.3 Définition des syndromes

À partir des trois symptômes précédents, nous définissons deux syndromes d'intérêt pour les cliniciens, utiles pour la formulation de diagnostics (Gauld *et coll.*, 2021).

Syndrome n°1 : perception subjective de somnolence excessive (PSSE)

Le syndrome de perception subjective de somnolence excessive (PSSE) reflète une plainte générale des sujets à propos d'une somnolence excessive. Nous le définissons par une KSS moyenne supérieure à 5 et un score à l'ESS supérieur à 15 (symptôme n°2 $\cap$ symptôme n°3). Pour cette tâche de classification binaire, 12 patients (12.9% du corpus) présentent ce syndrome.

Syndrome n°2 : propension sévère à l'endormissement diurne (PSED)

Le syndrome de propension sévère à l'endormissement diurne (PSED) mesure la tendance pathologique à l'endormissement, quand la plainte subjective de SDE (score à l'ESS supérieur à 15) est objectifiée par une latence d'endormissement au TILE inférieure à 8 minutes (symptôme n°1 $\cap$ symptôme n°2). Pour cette tâche de classification binaire, 12 patients (12.9%) présentent ce syndrome.

17.3 Système de classification

Le système de classification utilisé dans ce chapitre est identique à celui introduit dans le chapitre 16 précédent.

Cependant, une différence notable est la façon dont est faite la validation croisée. En effet, nous avons vu dans le chapitre précédent que la validation croisée double conduisait à des performances de classification relativement faibles, expliquée par le taille restreinte de la base de données et du faible nombre de locuteurs pour permettre à cette procédure de converger vers un classifieur robuste.

Nous décidons dans cette partie d'estimer les hyperparamètres de systèmes et de bloc au sein d'une seule validation croisée de type *Leave One Speaker Out Cross Validation* : à chaque itération, un locuteur est isolé afin de servir de test, et le système est entraîné sur les locuteurs restants. Même si cette pratique fait perdre au système sa capacité à généraliser sur d'autres données et augmente les risques de surestimation des performances obtenues (Cearns *et coll.*, 2019), le but poursuivi ici est différent du chapitre précédent : nous ne souhaitons pas concevoir un système de classification pour une future application en clinique, mais nous nous servons de la classification comme un support pour étudier notre hypothèse. La performance en elle-même est moins importante que la différence de performances à système et corpus constants.

De plus, puisque nous souhaitons comparer plusieurs méthodes pour estimer le statut de somnolence pathologique de chaque locuteur selon plusieurs axes, la LOSOCV permet exactement de fournir une estimation pour chaque locuteur indépendamment des autres, et sans risque de biais lié à l'appartenance à un certain bloc lors de la validation croisée à k blocs (cf. chapitre 16 précédent).

17.4 Détection de syndromes

Pour estimer les deux syndromes définis dans la section 17.2, nous proposons deux stratégies différentes.

17.4.1 Méthode n°1 : estimation des syndromes directement à partir des descripteurs vocaux

RÉF.	TÂCHE	SMOTE	ERR. STA	MW α	%PCA	CLASSIF.	UAR_{agg} LOSO
SYNDROMES							
I	PSSE	SMOTE	[Del, Ins, Sub]	1.0	80%	SVC : C = 1, k='poly', γ = 'auto', d=3	**82.6%**
II	PSED	Borderline	[Correct, Ins, Sub]	0.05	70%	SVC : C = 1, k='poly', γ = 'auto', d=4	**85.6%**

TABLEAU 17.1 – Meilleurs paramètres et performances des systèmes détectant les *syndromes* entraînés sur les descripteurs vocaux sous LOSOCV.

Une première façon d'estimer les syndromes à partir des caractéristiques vocales est d'entraîner des systèmes de classification dédiés sur des données annotées avec le statut du syndrome. Ainsi, nous avons annoté le corpus TILE avec les deux syndromes définis dans la section 17.2 et entraîné deux systèmes dédiés sur les caractéristiques vocales selon la procédure décrite dans la section précédente. Les paramètres et performances correspondants sont reportés dans la première partie tableau 17.1. Les matrices de confusion correspondantes sont affichées dans la figure 17.2.

Dans les deux tâches, les classifieurs atteignent des UAR supérieurs à 80% : la détection de la PSED (système II) atteint un UAR de 85.6% tandis que la détection de la PSSE (système I) est réalisée correctement avec un UAR de 82.6%.

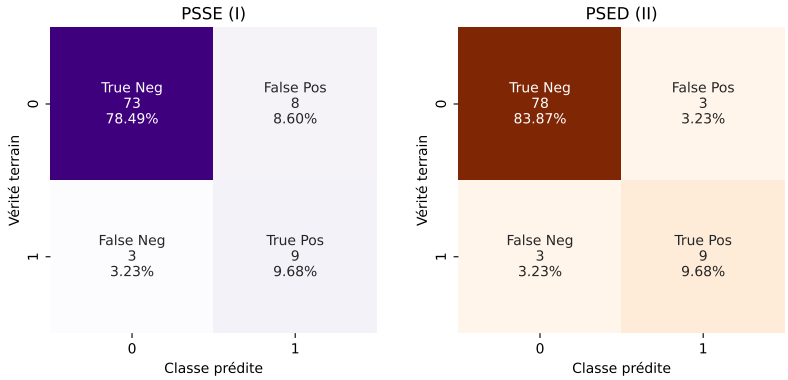

FIGURE 17.2 – Matrices de confusion des systèmes I et II estimant les syndromes directement à partir des marqueurs vocaux.

17.4.2 Méthode n°2 : Estimation des syndromes par la fusion des estimations des symptômes

Une deuxième façon d'estimer les syndromes à partir de marqueurs vocaux est d'estimer dans un premier temps les symptômes qui les constituent (cf section 17.2) et d'ensuite combiner les estimations obtenues, en appliquant l'opérateur logique ET – noté $\cap$ – aux classes binaires estimées, pour détecter le syndrome lui-même.

Estimation des symptômes

Nous avons rapporté dans le tableau 17.2 les performances des trois classifieurs entraînés et évalués sur les symptômes. Les matrices de confusion correspondantes sont affichées dans la figure 17.3.

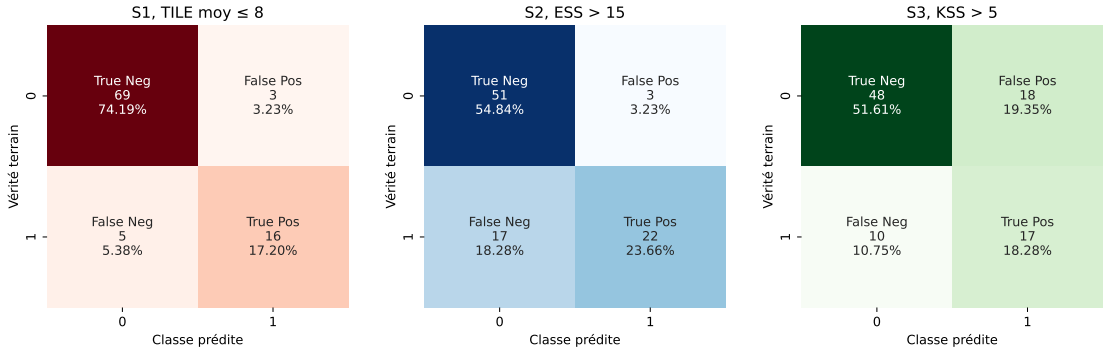

FIGURE 17.3 – Matrices de confusion correspondant aux systèmes S1, S2 et S3 pour la classification des symptômes à partir des marqueurs vocaux.

La meilleure performance est obtenue sur la classification du symptôme n°1 avec un UAR de 84.6%, tandis que l'équivalent subjectif (symptôme n°2) est détecté avec une précision de 75.4%. Enfin, le symptôme le plus difficile à détecter à partir des caractéristiques vocales avec notre pipeline est la somnolence moyenne (symptôme n°3), avec un UAR de 67.8%.

Par ailleurs, les trois systèmes intègrent un système de filtrage des erreurs de STA. Pour la détection de la propension à l'endormissement diurne mesuré objectivement (symptôme

RÉF.	TÂCHE	SMOTE	ERR. STA	MW α	%PCA	CLASSIF.	UAR_{agg} LOSO
SYMPTÔMES							
S1	TILE moy. ≤ 8	SMOTE	[Sub, Ins]	0.1	80%	SVC : C = 1, k='poly', γ = 'auto', d=4	**84.6%**
S2	ESS > 15	SMOTE	[Del, Ins, Sub]	-	70%	SVC : C = 1, k='poly', γ = 'auto', d=4	**75.4%**
S3	KSS moy. > 5	Borderline	[Ins, Sub]	0.075	85%	SVC : C = 1, k='poly', γ = 'auto', d=2	**67.8%**

TABLEAU 17.2 – Meilleurs paramètres et performances des systèmes estimant les *symptômes* entraînés sur les descripteurs vocaux sous LOSOCV.

n°1, TILE moy.), les erreurs sélectionnées sont les mêmes que dans le chapitre 14 (Insertions et Substitutions), alors que pour la perception subjective de SDE sévère (symptôme n°2, ESS), l'ensemble des erreurs sélectionnées est l'antagoniste de celui du chapitre précédent (Délétions, Insertions et Substitution vs Correct dans le chapitre précédent). Enfin, le système détectant la somnolence moyenne (symptôme n°3, KSS moy.) ne conserve que les Insertions et les Substitutions pour tous les systèmes, ce qui correspond aux erreurs des STA qui avaient été identifiées comme étant prédominantes dans le chapitre 14.

Une différence avec les précédents systèmes est malgré tout le classifieur, qui est pour les trois systèmes SVM avec un noyau polynomial de degré 2 ou 4. Contrairement au LDA et aux régressions linéaires employées dans le chapitre précédent, les SVM avec noyau ne sont pas interprétables et nous ne pourrons donc pas proposer d'étude approfondie des paramètres appliqués aux descripteurs par les classifieurs.

Estimation des syndromes

Les UAR obtenus par l'estimation de deux syndromes en utilisant la fusion des deux symptômes les composant sont représentés dans le tableau 17.3. Les matrices de confusion correspondantes sont affichées respectivement à gauche et au centre de la figure 17.4.

RÉF.	MÉTHODE		PSSE	PSED
(A)	Symptômes n°2 $\cap$ n°3	S2 $\cap$ S3	64.2%	-
		(I)	**82.6%**	-
(B)	Symptômes n°1 $\cap$ n°2	S1 $\cap$ S2	-	82.1%
		(II)	-	**85.6%**

TABLEAU 17.3 – UAR des différentes méthodes pour estimer les syndromes à partir des marqueurs vocaux ou des estimations des symptômes.

17.5 Discussion

17.5.1 Les symptômes sont-ils nécessaires pour estimer les syndromes ?

Pour l'estimation du syndrome de perception subjective de somnolence excessive (PSSE), les meilleures performances sont obtenues en entraînant un système de classification directement sur les marqueurs vocaux (I, UAR=82.6%). Calculer la fusion des estimations faites par les systèmes S2 et S3 conduit à un UAR significativement plus faible : puisque l'estimation

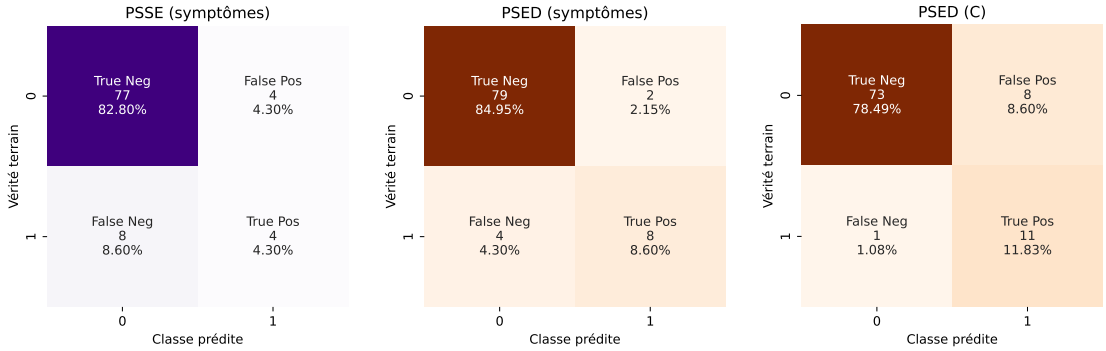

FIGURE 17.4 – Matrices de confusion des systèmes estimant les *syndromes* à partir de leurs *symptômes*.

du symptôme n°3 (KSS $>$ 5) – faisant partie de la définition du syndrome de PSSE – par le système S3 n'est pas aussi efficace que les autres systèmes d'estimation des symptômes (67.8% d'UAR vs plus de 75%), le fusion des estimations faites par les systèmes estimant les symptômes n°2 et n°3 n'en est pas performante.

Au contraire, l'estimation du syndrome de PSED à partir de la fusion tardive entre l'estimation des symptômes n°1 et n°2 (S1 $\cap$ S2) et l'estimation directe du syndrome à partir de la voix (II) aboutit à des performances du même ordre de grandeur (respectivement UAR=82.1% et UAR=85.6%).

17.5.2 L'estimation d'un seul symptôme est-elle suffisante pour estimer un syndrome ?

Puisque la fusion des estimations de deux symptômes donne des précisions équivalentes à celles de l'estimation directe pour l'estimation du syndrome de PSED, nous cherchons à savoir si l'un des deux symptômes définissant un syndrome est suffisant pour estimer le syndrome entier. Les UAR obtenus dans l'estimation de chacun des syndromes en utilisant un seul symptôme sont rapportés dans le tableau 17.4. Nous avons rapporté à la fois l'UAR obtenu avec la classe prédite de ce symptôme (S1, S2 et S3) et l'UAR attendu si son estimation était parfaite (dénoté par *oracle* dans le tableau 17.4).

RÉF.	MÉTHODE		PSSE	PSED
(C)	Symptôme n°1	*oracle*	-	94.4 %
	TILE $\leq$ 8	S1	-	**90.9%**
(D)	Symptôme n°2	*oracle*	83.3%	83.3%
	ESS $>$ 15	S2	58.5%	72.8%
(E)	Symptôme n°3	*oracle*	90.7%	-
	KSS $>$ 5	S3	61.9%	-

TABLEAU 17.4 – UAR des différentes méthodes pour estimer les syndromes à partir d'un seul symptôme.

Dans le cas du syndrome de PSSE, le symptôme n°3 (KSS $>$ 5) a de meilleures performances (UAR=61.9%) que le symptôme n°2 (UAR=58.5%), mais ces performances ne sont toujours pas suffisantes pour détecter le syndrome avec précision à partir de l'estimation d'uniquement un symptôme. L'UAR théorique qui peut être obtenu pour l'estimation de la PSSE à partir du symptôme n°3 est très élevé (UAR=90.7%), mais, encore une fois, les mau-

vaises performances du système S3 empêchent d'obtenir une bonne prédiction du syndrome PSSE à partir de l'unique estimation du symptôme n°3.

Cependant, dans le cas du syndrome de PSED, l'estimation à partir d'un seul des deux symptômes (symptôme n°1) est plus performante que l'approche précédente (UAR=90.9%, matrice de confusion à droite de la figure 17.4). Ce résultat – obtenu sur l'estimation du syndrome de PSED à partir de l'estimation du symptôme n°1 – est même supérieur à l'UAR obtenu dans l'estimation du symptôme lui-même par le système S1. Ce résultat peut indiquer que le symptôme n°1 est prédominant dans le syndrome de PSED et que le système S1, entraîné sur des données étiquetées avec ce symptôme, a généralisé à un niveau conceptuel plus élevé, apprenant à classifier la PSED au lieu du symptôme n°1.

17.6 Conclusion et perspectives

En conclusion, nous nous sommes rapprochés du raisonnement clinique en proposant la classification de deux syndromes liés à la somnolence, selon deux stratégies : soit directement à partir des marqueurs vocaux, soit en estimant d'abord les symptômes qui les composent.

Pour l'estimation du syndrome de perception subjective de somnolence excessive, l'estimation directe est plus performante que la fusion de l'estimation des deux symptômes, alors que pour l'estimation du syndrome de propension sévère à l'endormissement diurne, les deux approches sont similaires. Dans ce dernier syndrome, la meilleure méthode est cependant l'estimation de son symptôme dominant, c'est-à-dire la propension objective au sommeil.

Au regard du nombre de symptômes mesurés dans la base TILE, de nombreux autres syndromes auraient pu être définis, par exemple sur la base de l'index de somnolence de Barcelone (BSI) ou du test d'alerte de l'hôpital de Toronto (THAT). La définition de ces différents syndromes reste malgré tout limitée par la population incluse dans la base de données, qui possède déjà des caractéristiques spécifiques (patients du pôle universitaire de médecine du sommeil du CHU de Bordeaux).

En continuant dans la direction impulsée par ce chapitre, nous proposons dans le chapitre suivant une réflexion plus approfondie sur le rôle des symptômes dans la détection de pathologies à travers des marqueurs vocaux, en présentant une approche novatrice basée sur les réseaux de symptômes.

Chapitre 18

Du *syndrome* aux *réseaux de symptômes*

Sommaire

Ce chapitre reprend et complète les réflexions commencées avec Jean-Arthur Micoulaud-Franchi (MCU-PH Université de Bordeaux, CHU de Bordeaux) et Christophe Gauld (CCA, CHU de Lyon et doctorant en philosophie des sciences, Université Paris 1 Panthéon-Sorbonne) lors de l'élaboration de mon mémoire pour le Diplôme Inter-Université de Philosophie et d'Épistémologie de la psychiatrie, organisé par les universités de Bordeaux, Toulouse et Marseille. La dernière section de ce chapitre est largement inspirée et basée sur l'ouvrage de Christophe Gauld, *Les réseaux de symptômes en psychopathologie : Enjeux théoriques, méthodologiques et sémiologiques* (Gauld, 2021).

18.1 Contexte et motivation

Ce chapitre poursuit la réflexion du chapitre précédent sur la translation d'un problème clinique en tâche pour l'apprentissage automatique. Dans la section 18.2, nous proposons d'étudier, à travers l'exemple de la détection de la dépression dans la voix, qui est une tâche occupant une grande part de recherche dans le domaine de la détection de pathologies dans la voix (cf. chapitre 2), les limites de la formulation du problème de la détection de pathologies grâce à des marqueurs vocaux (partie A de la figure 18.1).

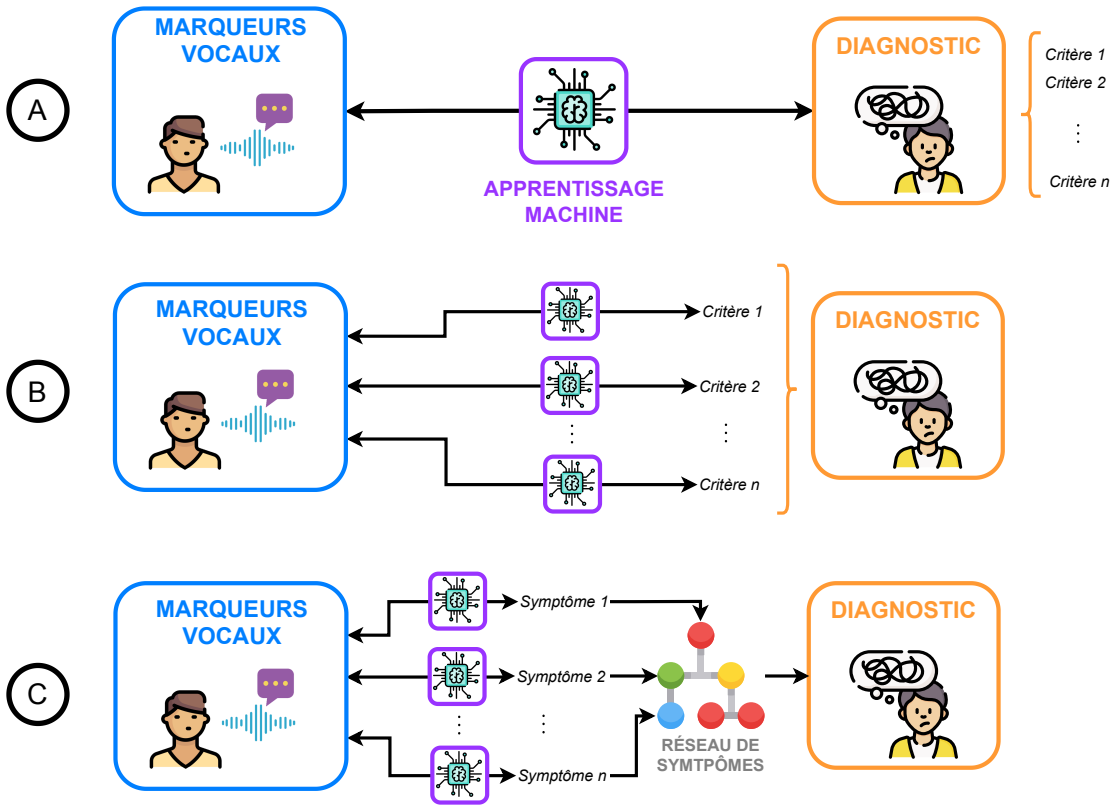

FIGURE 18.1 – Trois différentes approches pour détecter une pathologie grâce à des marqueurs vocaux. (A) Estimation directe de la pathologie. (B) Estimation des symptômes composant les critères diagnostiques. (C) Estimation en utilisant un réseau de symptômes.

Dans la section 18.3, nous présentons les avantages de formuler les problèmes d'apprentissage automatique en médecine sous la forme d'estimation de symptômes plutôt qu'une estimation directe de la pathologie. Enfin, nous proposons dans la section 18.4 une approche

récente pour formuler différemment, à la fois en informatique et en médecine, le lien entre symptômes, pathologies et mesures écologiques (dont la voix) : les réseaux de symptômes.

18.2 L'exemple de la détection de la dépression dans la voix

Lors de la revue proposée dans le chapitre 2, nous avons pu identifier deux différentes façons d'annoter la dépression dans les corpus utilisés :

— soit le score à un questionnaire de dépression, comme le questionnaire de santé des patients [PHQ-8 (Spitzer *et coll.*, 1999; Kroenke *et coll.*, 2009)] ou l'inventaire de dépression de Beck [BDI-II, (Beck, 1996)], dont les questions évaluent différentes dimensions de la pathologie. Un seuil est ensuite appliqué (10 pour la PHQ-8 et 18 pour le BDI-II) pour classifier les sujets entre « atteint de dépression » et « non atteint de dépression ». C'est le cas par exemple du DAIC-WOZ et du corpus AVEC 2013, qui sont les corpus les plus utilisés sur cette tâche.

— soit le diagnostic par des médecins, qui sur la base d'un critère diagnostic (généralement le DSM-IV ou 5, parfois non précisé) vont diagnostiquer le patient comme « atteint de dépression » ou « non atteint de dépression ». Pour notre propos, nous mettons en avant la base de donnée présentée dans Di *et coll.* (2021), qui est la plus grosse base de données sur la détection de la dépression dans la voix à notre connaissance (3580 patients « atteints de dépression »/4016 sujets « non atteints de dépression »).

Dans les deux sections suivantes, nous mettons en lumière les contradictions de ces deux pratiques avec un usage réel et efficace des systèmes conçus sur ces corpus.

18.2.1 Score à un questionnaire global : l'exemple de la PHQ-8

Dans cette section, nous prenons l'exemple de la PHQ-8 pour illustrer les limites de l'annotation de bases de données avec des questionnaires de pathologie. Ce questionnaire inclut huit symptômes : les changements d'appétit, le sentiment d'échec ou d'inutilité, la léthargie, les problèmes de sommeil, les problèmes de concentration, la perte d'intérêt et de capacité à prendre du plaisir, un sentiment déprimé, une altération des capacités psychomotrices.

L'utilisation d'autoquestionnaires bénéficie de plusieurs avantages. Tout d'abord, ceux-ci coûtent peu et sont faciles à mettre en place. Ils requièrent souvent très peu de temps pour les remplir, et peuvent être administrés par n'importe quelle personne, puisque c'est le patient lui-même qui s'évalue.

Cependant, cette approche souffre d'un défaut de sensibilité à la pathologie. En effet, dans la méta-analyse menée sur 17 études cliniques par Gilbody *et coll.* (2007), la PHQ-8 a démontré une sensibilité de seulement 77% comparé aux mêmes sujets diagnostiqués par des professionnels. Ainsi, même un algorithme hypothétique permettant de reproduire parfaitement la base de données sur lequel il est entraîné (100% de performances) ne pourra atteindre qu'au mieux 77% de sensibilité dans la classification de la dépression. Cette approche induit donc une cascade d'approximations (approximation due à l'utilisation du questionnaire et approximation due à l'algorithme de classification), qui contribue à la divergence entre la modélisation numérique de la dépression et sa réalité (Arseniev-Koehler *et coll.*, 2018).

Les limites des questionnaires pour le diagnostic de pathologies et la nécessité de l'entretien clinique pour le diagnostic trouvent leur exemple extrême dans le diagnostic de la schizophrénie. En effet, aucun des symptômes qui guident le diagnostic de cette pathologie ne lui est spécifique : « Il est rare pour un clinicien d'être capable de décrire précisément le processus diagnostic de la schizophrénie. » (Rümke, 1958) [cité dans la thèse de médecine

de Serafino (2020)]. Les cliniciens utilisent alors dans ce cas leur *praecox feeling*, sentiment de conviction du clinicien, étudié en profondeur dans (Serafino, 2020).

Par ailleurs, les questionnaires ne sont globalement pas utilisés par les cliniciens (Zimmerman et McGlinchey, 2008) : un système d'aide au diagnostic estimant la probabilité qu'un patient soit au-dessus (ou en dessous) d'un seuil pathologique sur un questionnaire, comme cela est le cas dans la majorité de la littérature n'a que peu d'intérêt pour le clinicien, qui est discutée dans la prochaine section.

Enfin, la validité des questionnaires médicaux dépend directement des études de validation et de réplicabilité qui ont été faites pour leur validation : la fiabilité des questionnaires comme vérité terrain pour un corpus dépend ainsi directement de la fiabilité du diagnostic des sujets de chaque groupe.

18.2.2 Diagnostic par un clinicien

Une meilleure approche semble donc le diagnostic par un clinicien, qui, sur la base d'une classification de référence (en majorité le DSM-IV ou le DSM 5 dans les corpus mentionnés dans le chapitre 2), diagnostique le sujet comme « atteint de dépression » ou non. Cette approche répond aux objections soulevées dans la partie précédente, puisque le diagnostic est directement reporté dans la base de données, sans perte de sensibilité due aux questionnaires.

L'inconvénient de cette approche est la mesure des critères diagnostiques inclus dans la classification. Les 7 échelles les plus utilisées produisent à elles seules 52 façons de mesurer ces critères (Fried, 2017). Si l'on combine cette multiplicité de mesures avec les critères diagnostiques du DSM-5 pour l'épisode de dépression majeure, pour lequel il faut au moins 5 symptômes sur les neuf incluant au moins l'humeur triste et/ou l'anhédonie, cela conduit à 10377 profils sémiologiques uniques. Ce chiffre n'est pas que théorique puisque, dans l'étude expérimentale proposée par (Fried et Nesse, 2015), 1030 profils sémiologiques ont pu être extraits sur un total de 3703 patients.

Ainsi, même dans la base de données présentée par Di *et coll.* (2021) qui contient pourtant 7596 sujets, les différences intergroupes sont noyées dans les différences interindividuelles. Cette approche n'est donc pas adaptée pour les pathologies qui ont de fortes prévalences et/ou une hétérogénéité interne forte, dont la dépression fait partie.

Par ailleurs, les critères diagnostiques d'une pathologie peuvent être amenés à changer, comme en témoignent les différentes versions du DSM : comment garantir la pérennité de l'effort investi dans l'annotation de base de données et dans l'entraînement de modèles basés sur celles-ci si cette annotation a une durée de vie potentiellement limitée ?

Enfin, face à certaines estimations allant jusqu'à 50% de la population générale rentrant dans le critère diagnostique d'une pathologie psychiatrique, une crise du diagnostic en psychiatrie a emergé dans les années 2000/2010 (Kirmayer *et coll.*, 2015), questionnant de nouveau de manière profonde la notion du normal et du pathologique (Frances, 2013). Même si le diagnostic fait par les cliniciens ne se résume pas à un algorithme appliquant de manière automatique un ensemble de critères diagnostiques (Blashfield et Herkov, 1996), l'étude précédente sur l'homogénéité de la définition de la dépression jette de sérieux doutes sur la fiabilité d'une telle pratique.

18.3 Annotation des symptômes

Nous proposons dans cette partie quelques avantages en faveur de l'annotation des bases de données avec les *symptômes* plutôt qu'avec le statut diagnostique des sujets.

18.3.1 Unité élémentaire de la sémiologie

Comme précisé dans le chapitre 17 précédent, les symptômes et traits sont l'unité la plus élémentaire du raisonnement clinique, à partir de laquelle sont élaborés les syndromes et les diagnostics. À partir des symptômes inclus dans le corpus, il est donc théoriquement possible de reconstruire les syndromes – voire les critères diagnostiques – à partir desquels les médecins posent leur diagnostic, garantissant une continuité avec les travaux précédents utilisant des bases de données annotées directement avec un diagnostic (partie B de la figure 18.1).

18.3.2 Explication mécanistique

De plus travailler au niveau des symptômes plutôt qu'à celui de la pathologie favorise l'explication mécanistique (Norman, 2000) : il semble moins complexe de relier un marqueur vocal à une autre manifestation physique ou psychique qu'à une pathologie, dont les mécanismes peuvent ne pas être connus.

Par ailleurs, au regard des résultats présentés dans le chapitre 17 précédent – et notamment le fait qu'un des deux syndromes est mieux estimé par un système entraîné sur son symptôme dominant, une relecture plus fine de la revue proposée dans le chapitre 2 permettrait peut-être de mettre au jour des systèmes estimant, à l'inverse, un symptôme dominant plutôt que le syndrome en lui-même. Par exemple, pour le trouble dépressif majeur, le symptôme central habituellement estimé est la tristesse de l'humeur (par ex. (Schultebraucks *et coll.*, 2020; Shinohara *et coll.*, 2021)). Cependant, au regard des cohortes et des paradigmes expérimentaux utilisés (patients diagnostiqués vs sujets sains, sans contrôle de l'humeur dans le groupe contrôle), il est difficile de préciser le niveau conceptuel que le classifieur a généralisé, et donc de s'assurer qu'il classifie bien la pathologie.

18.3.3 Nécessaires au diagnostic différentiel et au pronostic

Par ailleurs, les symptômes sont indispensables au diagnostic différentiel et au pronostic des pathologies étudiées, tâches actuellement très peu représentées dans les travaux de la communauté du traitement du signal de voix.

18.3.4 Constance dans le temps

Les symptômes bénéficient également d'une constance historique : les plaintes exprimées par les patients sont les mêmes, quelles que soient les époques [par ex. (Magiorkinis *et coll.*, 2009) pour le mal de tête ou (Zarate et Manji, 2009) pour les troubles de l'humeur]. Au contraire, comme vu précédemment, les classifications ont des durées de vie limitées, dépendant de leur mise à jour avec les nouvelles connaissances accumulées entre les différentes versions.

18.3.5 Différences culturelles des diagnostics

Par ailleurs, annoter les bases de données directement avec les diagnostics ne prend pas en compte les différences culturelles qui peuvent exister entre les classifications, et la prise en compte de l'aspect social de la définition même des maladies (Kirmayer, 1989, 1991; Westermeyer, 1985).

Dans son rapport sur l'éthique et la gouvernance de l'IA pour la santé de 2021 (World Health Organization, 2021), l'Organisation mondiale de la santé (OMS) a mis en garde contre les éventuels biais – notamment culturels – que peuvent avoir les algorithmes mal entraînés

et les conséquences que ceux-ci peuvent avoir sur la santé des populations impliquées : une annotation avec les symptômes permet de prendre en compte ces différences socioculturelles et contribue ainsi à des systèmes d'IA plus éthiques.

18.3.6 Efforts de recherche mieux distribués

Enfin, travailler directement sur les symptômes permet de mutualiser efficacement les efforts de recherches à un niveau global. En lieu et place de nombreuses équipes travaillant en parallèle sur des pathologies distinctes à partir de petites bases de données, ne donnant que des résultats épars et n'ayant aucune application clinique possible, la répartition de l'effort de recherche sur les symptômes permettrait de mettre en commun les résultats obtenus pour diagnostiquer de nombreuses maladies.

18.4 Réseaux de symptômes

18.4.1 Présentation des réseaux de symptômes

Plutôt qu'organiser les symptômes au sein d'une liste comme c'est le cas dans les classifications usuelles, une approche proposée par Borsboom consiste à les organiser dans une structure de graphe, appelée *réseaux de symptômes* (Borsboom et Cramer, 2013).

Chaque noeud du réseau représente un symptôme – voire dans certains modèles des mesures biologiques ou environnementales – et les liens entre noeuds correspondent à la co-occurrence des symptômes dans une population cible. Ainsi, deux symptômes qui sont présents simultanément dans la majorité des exemples de la base de données d'entraînement auront un lien plus fort que deux symptômes n'apparaissant jamais en même temps. Nous ne traiterons dans cette partie que les réseaux de symptômes dont les arêtes ne sont pas dirigées (pas de sens favorisé de l'interaction entre symptômes), mais certains modèles utilisent des graphes dirigés ou partiellement dirigés pour expliciter des relations de causalité déjà identifiées (Gauld, 2021).

Un exemple fictif inspiré des réseaux de symptômes proposés dans (McElroy *et coll.*, 2019) pour les épisodes dépressifs majeurs chez les enfants et adolescents est représenté dans la figure 18.2.

18.4.2 Définition du pathologique

À partir de ce graphe, un état pathologique est défini comme un état stable du réseau pour lequel des symptômes sont observés. Nous proposons dans la figure 18.3 le processus de passage d'un réseau de symptôme de son état stable non pathologique (symptômes latents) à son état pathologique (symptômes observés), à partir d'un évènement extérieur qui active deux symptômes du réseau.

Ce phénomène se fait selon la chronologie suivante :

1. Un élément extérieur E1 survient, potentiellement capable d'activer les deux symptômes « Tristesse de l'humeur » et « Perte d'estime de soi » ;
2. Les deux symptômes sont activés ;
3. Le réseau, qui est en déséquilibre, active les symptômes qui sont les plus liés aux deux symptômes précédemment activés, jusqu'à retrouver un état d'équilibre. En fonction des caractéristiques du réseau, cet équilibre peut se faire même quand l'évènement extérieur n'est plus présent ;

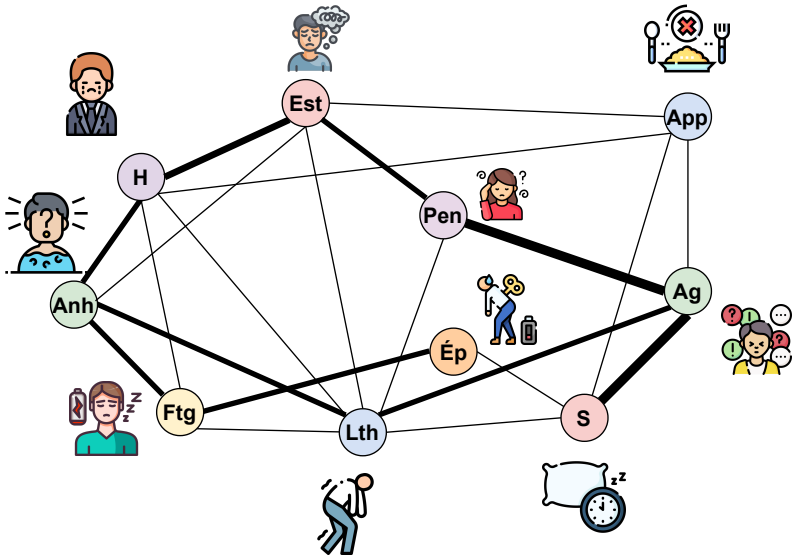

FIGURE 18.2 – *H* : Tristesse de l'humeur, *Anh* : Anhédonie, *Ftg* : Fatigue, *Lth* : Léthargie, *Ép* : Épuisement, *S* : Troubles du sommeil, *Ag* : Agitation, *Pen* : Problèmes pour penser clairement, *App* : Troubles de l'appétit, *Est* : Perte d'estime de soi [exemple fictif inspiré de (McElroy *et coll.*, 2019)].

4. Le réseau retrouve un état stable (ici avec les symptômes activés – donc dans un état stable pathologique).

L'état final du réseau de symptômes dépend du ou des symptômes qui ont été initialement activés, mais aussi de la connectivité du réseau.

18.4.3 Perspectives de recherche autour des réseaux de symptômes

Ce modèle propose ainsi une autre façon d'organiser les symptômes liés à une pathologie. En réutilisant les critères diagnostiques d'une classification sous forme de réseau de symptômes, nous obtenons un modèle *quantitatif* des relations entre symptômes qui sont décrites de manière *qualitative* dans les classifications. Nous proposons dans cette section quatre perspectives sur l'utilisation des marqueurs vocaux dans les réseaux de symptômes.

Multimodalité

Un grand avantage des réseaux de symptômes est la facilitation de l'implémentation d'une détection multimodale des pathologies modélisées, dans une visée de médecine numérique écologique. Alors que les systèmes de détection de pathologies tendent de plus en plus à devenir multimodaux (Nie *et coll.*, 2015), tous les symptômes ne sont pas nécessairement estimés de manière optimale par un unique médium de mesure. Avec les réseaux de symptômes, un symptôme peut être estimé correctement avec des marqueurs vocaux tandis qu'un autre pourrait l'être par le nombre de pas ou des caractéristiques d'utilisation de smartphone [par ex. (Low *et coll.*, 2017)], l'étape de *fusion* de l'information se faisant directement dans le réseau.

Granularité adaptative

Un deuxième avantage de recourir aux réseaux de symptômes en médecine numérique est la *granularité adaptative* qu'elle permet. Alors que de plus en plus de bases de données

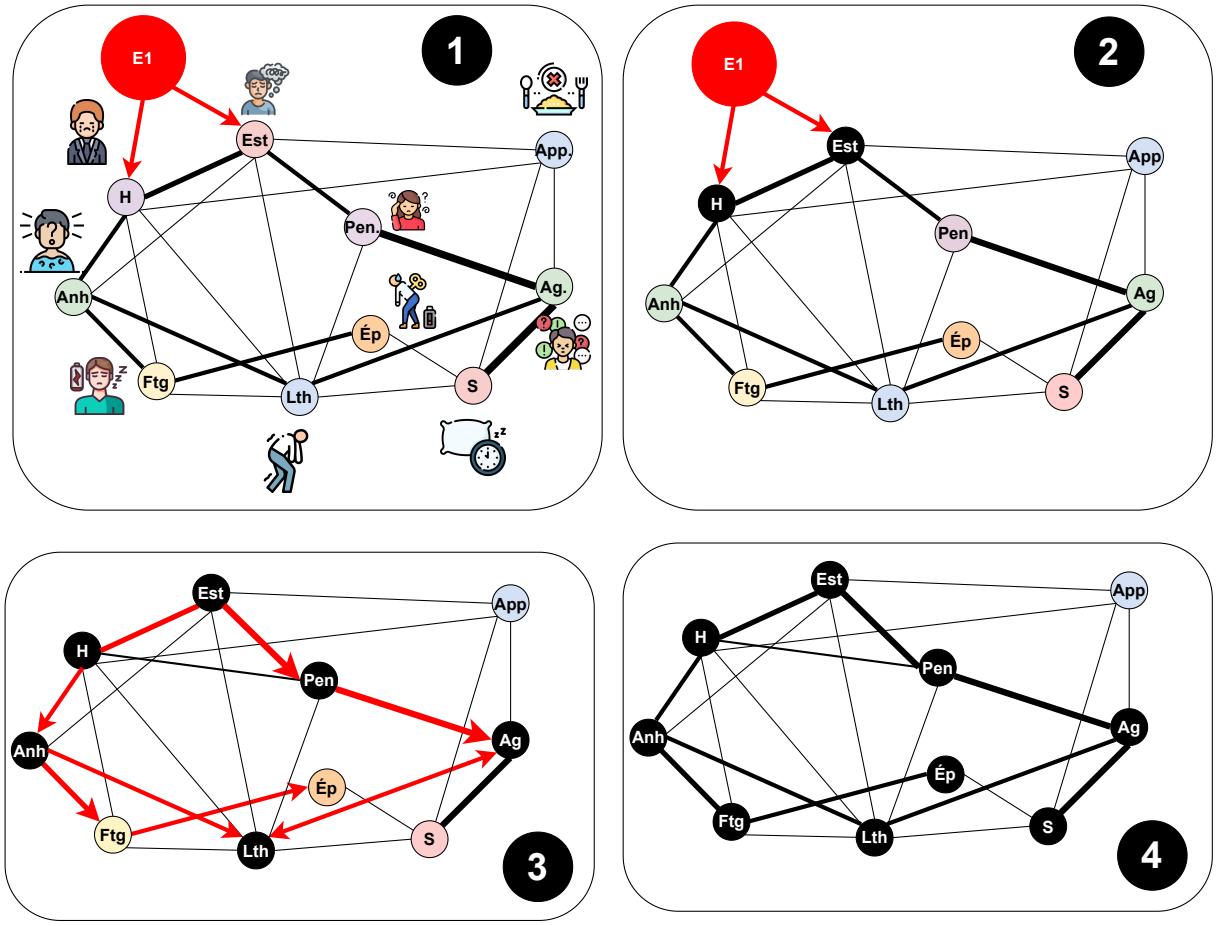

FIGURE 18.3 – Exemple d'activation successive des symptômes d'un état stable sain à un état stable pathologique suite à un évènement extérieur E1 activant deux symptômes. Les symptômes activés sont représentés en noir.
H : Tristesse de l'humeur, *Anh* : Anhédonie, *Ftg* : Fatigue, *Lth* : Léthargie, *Ép* : Épuisement, *S* : Troubles du sommeil, *Ag* : Agitation, *Pen* : Problèmes pour penser clairement, *App* : Troubles de l'appétit, *Est* : Perte d'estime de soi [exemple fictif inspiré de (McElroy *et coll.*, 2019)].

tendent à être collectées en conditions écologiques, par les patients eux-mêmes, afin de créer de grandes cohortes, la liste de symptômes à mesurer peut changer au cours de la collecte de données. Les réseaux de symptômes permettent de changer en cours de collecte les symptômes collectés, sans pour autant perdre l'information apprise avec un symptôme d'une autre granularité.

En effet, les réseaux de symptômes permettent, si nécessaire, de remplacer un noeud (représentant un symptôme) par un sous réseau de granularité plus fine. Dans la figure 18.4, nous reprenons l'exemple de réseau précédent, en affinant le symptôme « Troubles du sommeil » par un sous réseau comprenant, par exemple, l'insomnie, l'inertie au sommeil, les parasomnies et la somnolence diurne excessive. Grâce à des algorithmes sur les graphes (comme par exemple l'algorithme de *propagation de croyances généralisé*, voir dernier paragraphe de cette section), l'information apprise sur les liens entre troubles du sommeil et les symptômes qui y sont liés ne sont pas perdus lors du remplacement de l'unique noeud par un sous-réseau.

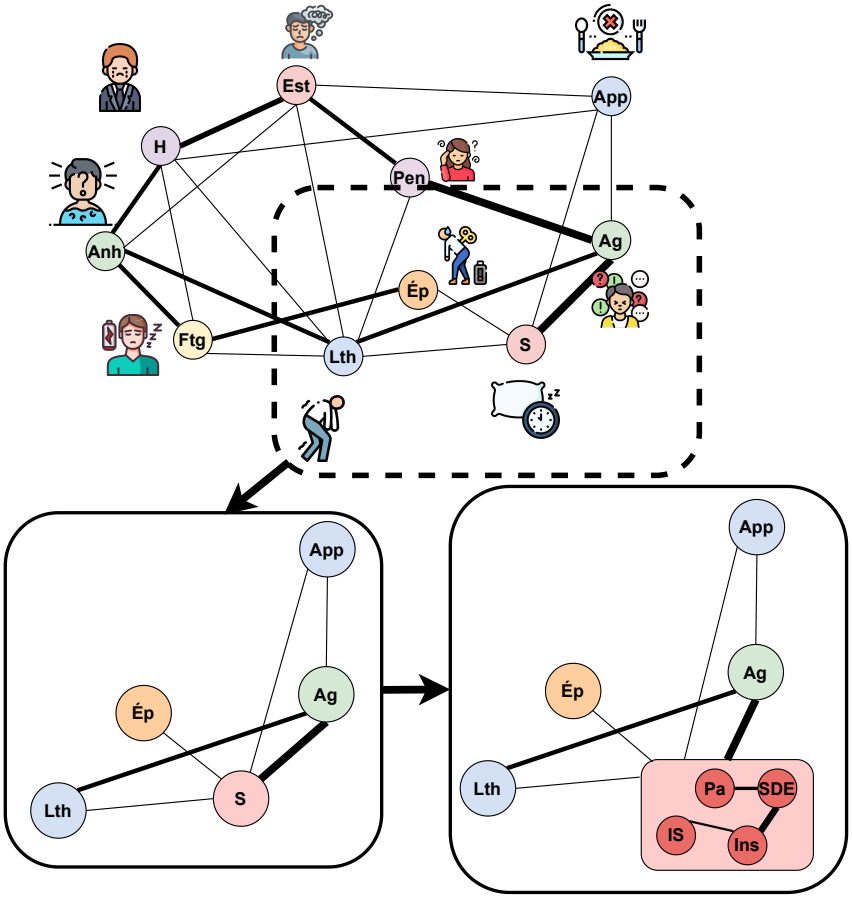

FIGURE 18.4 – Adaptation de granularité pour les troubles du sommeil, qui sont remplacés dans le réseau par un sous réseau comprenant l'inertie au sommeil (*IS*), les parasomnies (*Pa*), la somnolence diurne excessive (*SDE*) et les insomnies (*Ins*).
H : Tristesse de l'humeur, *Anh* : Anhédonie, *Ftg* : Fatigue, *Lth* : Léthargie, *Ép* : Épuisement, *S* : Troubles du sommeil, *Ag* : Agitation, *Pen* : Problèmes pour penser clairement, *App* : Troubles de l'appétit, *Est* : Perte d'estime de soi [exemple fictif inspiré de (McElroy *et coll.*, 2019)].

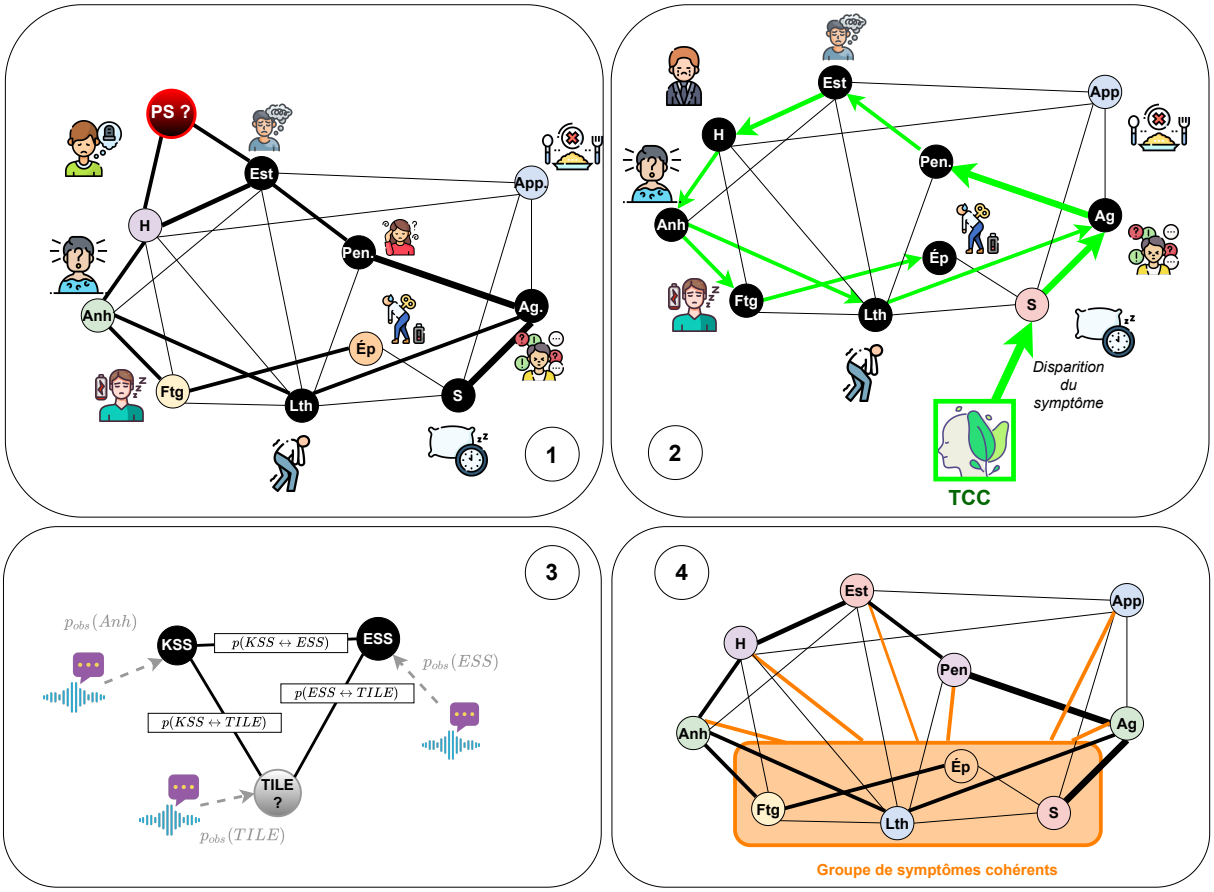

FIGURE 18.5 – Perspectives de recherches apportées par les réseaux de symptômes. 1) Estimation d'un symptôme inaccessible (exemple des pensées suicidaires). 2) Ciblage thérapeutique du symptôme qui désactivera les symptômes du réseau. 3) Affinage des estimations des états de chaque symptôme par la voix (ici la latence d'endormissement moyenne pathologique au TILE) en utilisant le graphe sous-jacent au réseau de symptômes. 4) Utilisation de l'algorithme du *generalized belief propagation* pour ajouter, en plus des connexions entre symptômes, des connexions entre groupes de symptômes. *PS* : pensées suicidaires, $p(symptôme1 \leftrightarrow symptôme2)$: probabilité de co-occurence des symptômes $p_{obs}(symptôme)$: pseudoprobabilité du symptôme déterminé à partir de marqueurs vocaux.

Algorithmes de propagation de croyances

Enfin, recourir aux réseaux de symptômes permet de tirer parti de toute la littérature de recherche portant sur les graphes. Un exemple, que nous avons déjà utilisé pour la détection d'accords musicaux dans un fichier audio (Martin *et coll.*, 2019), est l'algorithme de *propagation de croyance – belief propagation* (Yedidia *et coll.*, 2001, BP). Cet algorithme permet, à partir des probabilités de transition entre les noeuds et de la probabilité a priori de chaque noeud, de trouver efficacement l'état d'équilibre d'un réseau entier ayant été précédemment déséquilibré.

Nous proposons quatre perspectives de recherches, d'intérêt à la fois pour la recherche clinique et pour la recherche en traitement du signal vocal. Celles-ci sont illustrées dans la figure 18.5.

Application n°1 : Estimation d'un symptôme inaccessible Un exemple de symptôme inaccessible est la présence de pensées suicidaires, qui est très difficile à estimer de manière directe

pour le clinicien (Yigletu *et coll.*, 2004). Utiliser un réseau de symptômes et un algorithme de BP permettrait d'estimer, à partir d'estimations des symptômes qui sont les plus liés à l'idéation suicidaire et du réseau tout entier, la probabilité de présence d'idées suicidaires chez un patient.

Application n°2 : Ciblage thérapeutique De nombreuses thérapies cognitivo-comportementales permettent de réduire une grande diversité de symptômes (Hofmann *et coll.*, 2012), mais il n'est pas toujours facile pour le clinicien de choisir quel symptôme diminuer en priorité. En simulant la réduction ou la disparition d'un symptôme dans un réseau et en appliquant un algorithme de BP pour trouver l'état stable final lié à la disparition du noeud correspondant, il est alors possible de cibler le ou les symptômes à traiter en priorité pour faire sortir le patient de son état stable pathologique.

Application n°3 : Robustesse des estimations de symptômes en apprentissage automatique Dans les chapitres 16 et 17, nous avons estimé de manière indépendante trois symptômes à partir de marqueurs vocaux. Il y aurait pourtant avantage à estimer un des symptômes en connaissance de l'estimation des deux autres (par ex. estimer une latence moyenne pathologique au TILE à partir de descripteurs vocaux et des estimations des niveaux de plainte de SDE et de somnolence moyenne). Pour cela construire un réseau dont les probabilités a priori des noeuds sont les pseudoprobabilités de présenter le symptôme (estimées par les classifieurs) et les probabilités de transition apprises sur le réseau de symptômes correspondant permettrait, après avoir trouvé l'état stable final du réseau, une correction des pseudoprobabilités et donc, au final, une meilleure estimation de chacun d'entre eux.

Application n°4 : Propagation de croyances généralisée Enfin, une dernière application notable de l'utilisation de l'algorithme de propagation de croyances dans les réseaux de symptômes est sa généralisation appelée propagation de croyance généralisée – *generalized belief propagation (GBP)*. Cet algorithme généralise le BP à des groupes des noeuds plutôt qu'aux noeuds eux-même (Yedidia *et coll.*, 2004; Sibel *et coll.*, 2012). Cet algorithme permet à la fois des perspectives de recherche clinique, en groupant des facteurs de comorbidités et en enrichissant le graphe avec des groupes de symptômes cohérents ; mais aussi des perspectives dans la construction des bases de données dynamiques, permettant une granularité variable des symptômes inclus dans le graphe (cf l'exemple précédent sur les troubles du sommeil, figure 18.4).

18.5 Conclusion

En conclusion, nous avons montré dans ce chapitre à travers l'exemple de la détection de la somnolence dans la voix les limites de l'annotation des bases de données directement avec le diagnostic des sujets, que ce soit par le biais de questionnaires ou du diagnostic par un spécialiste.

Nous avons présenté les avantages de travailler à l'échelle des symptômes, qui sont l'unité élémentaire de la sémiologie, indépendants de la culture des patients, et constants dans le temps.

Par ailleurs, cette approche ouvre des perspectives sur les réseaux de symptômes en médecine numérique, qui ont de nombreuses applications à la fois en recherche clinique et pour la construction de bases de données dynamiques.

Conclusion de la partie

Dans cette partie, nous avons proposé une réflexion sur la formulation du problème clinique de la détection de la somnolence dans la voix en termes d'apprentissage automatique.

Classification de symptômes

Dans un premier temps, nous avons entraîné trois classifieurs de symptômes dans une perspective d'implémentation en conditions réelles (*MLOps*). Pour chacun des symptômes étudiés – propension à l'endormissement diurne, somnolence diurne excessive sévère, et somnolence moyenne – les classificateurs sont identiques et incluent un bloc de décorrélation des descripteurs vocaux avec des cofacteurs pouvant influencer la voix et/ou la somnolence, permettant de garantir de la spécificité des marqueurs utilisés par les classifieurs vis-à-vis de ces cofacteurs. Après avoir détaillé la procédure d'entraînement de ces classificateurs, nous avons proposé une analyse des résultats et des descripteurs employés par ces classifieurs.

Aucun des symptômes étudiés n'a été classifié avec des performances supérieures à 70%, ce que nous attribuons au trop faible effectif de la base TILE, la première phase d'entraînement des classificateurs nécessitant deux boucles de validation croisées imbriquées l'une dans l'autre.

Classification de syndromes

Dans un deuxième temps, afin de nous rapprocher de la réalité du raisonnement clinique lors du diagnostic, nous avons étudié la classification de deux syndromes dérivés des trois symptômes précédents : la perception subjective de somnolence excessive (PSSE, définie par une ESS > 15 et une KSS moyenne > 5), et la propension sévère à l'endormissement diurne (PSDE, définie par une ESS > 15 et une latence d'endormissement au TILE ≤ 8 min.) Deux méthodes ont été proposées : d'une part, l'estimation directe de ces syndromes à partir des descripteurs vocaux ; d'autre part leur détection à partir d'une estimation des symptômes qui les compose.

Si la PSSE est la mieux classifiée directement à partir de la voix (82.6% d'UAR), les deux méthodes fournissent des performances de classification similaires pour la PSDE (resp. 85.6% et 82.1% d'UAR). Enfin, nous avons envisagé qu'un seul symptôme puisse être suffisant pour classifier ces syndromes : alors que cette approche ne fournit pas de résultat probant pour la PSSE (en raison de la faible estimation de la somnolence moyenne), la PSDE est correctement identifiée avec une performance de 90.9% uniquement grâce à l'estimation de la latence d'endormissement pathologique au TILE.

Réseaux de symptômes

Enfin, nous avons montré les limites des bases de données annotées directement avec le diagnostic d'une pathologie (par des questionnaires ou directement par les cliniciens), et montré les avantages d'une annotation basée sur les symptômes : ces derniers sont indépendants des cultures, stables dans le temps, facilitent l'explication mécanistique de leur expression à travers la voix et sont mutualisables pour l'étude de nombreuses pathologies. Par ailleurs,

cette approche permet l'utilisation de réseaux de symptômes, promettant des perspectives dans les cadres à la fois de la pratique clinique (inférence de symptômes inaccessibles, ciblage thérapeutique), de la collecte de bases de données dynamiques (granularité variable des symptômes inclus dans le réseau) et de l'estimation de ces symptômes (estimation conjointe de plusieurs symptômes).

Perspectives

Cette première réflexion sur la traduction d'une problématique clinique – la détection de la somnolence dans la voix – en problème d'apprentissage automatique – la classification binaire de symptômes ou de syndrome à partir de descripteurs vocaux – nécessiterait d'être approfondie : alors que la majorité des travaux produits par le communauté de traitement du signal de parole étudie la classification binaire d'un diagnostic, la traduction d'autres problématiques cliniques (et notamment du diagnostic différentiel et du pronostic) est nécessaire à la pratique clinique numérique de demain.

Un modèle qui nous semble particulièrement prometteur pour favoriser l'interaction entre les deux domaines impliqués dans la médecine numérique – médecine et traitement de données – est le réseau de symptômes, permettant de formaliser les problématiques cliniques à l'aide de structures connues des informaticiens, favorisant ainsi la recherche interdisciplinaire.

Bibliographie de la partie

Anguita, D., Ghelardoni, L., Ghio, A., Oneto, L., et Ridella, S. (**2012**). "The 'K'in K-fold cross validation," dans *20th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, i6doc. com publ, pp. 441–446.

Arand, D., Bonnet, M., Hurwitz, T., Mitler, M., Rosa, R., et Sangal, R. B. (**2005**). "The Clinical Use of the MSLT and MWT," SLEEP **28**(1), 123–144, doi: 10.1093/sleep/28.1.123.

Arseniev-Koehler, A., Mozgai, S., et Scherer, S. (**2018**). "What type of happiness are you looking for? - A closer look at detecting mental health from language," dans *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology : From Keyboard to Clinic*, Association for Computational Linguistics, New Orleans, LA, pp. 1–12, doi: 10.18653/v1/W18-0601.

Beck, A. T. (**1996**). *Beck depression inventory (BDI-II)*, **10**.

Bengio, Y., et Grandvalet, Y. (**2003**). "No Unbiased Estimator of the Variance of K-Fold Cross-Validation," dans *Advances in Neural Information Processing Systems*, édité par S. Thrun, L. Saul, et B. Schölkopf, MIT Press, Vol. 16.

Blashfield, R. K., et Herkov, M. J. (**1996**). "Investigating Clinician Adherence to Diagnosis by Criteria : A Replication of Morey and Ochoa (1989)," Journal of Personality Disorders **10**(3), 219–228, doi: 10.1521/pedi.1996.10.3.219.

Borsboom, D., et Cramer, A. O. (**2013**). "Network Analysis : An Integrative Approach to the Structure of Psychopathology," Annual Review of Clinical Psychology **9**(1), 91–121, doi: 10.1146/annurev-clinpsy-050212-185608.

Bostic, J. Q., et Rho, Y. (**2006**). "Target-Symptom Psychopharmacology : Between the Forest and the Trees," Child and Adolescent Psychiatric Clinics of North America **15**(1), 289–302, doi: 10.1016/j.chc.2005.08.003.

Bowen, J. L. (**2006**). "Educational strategies to promote clinical diagnostic reasoning," New England Journal of Medicine **355**(21), 2217–2225.

Cearns, M., Hahn, T., et Baune, B. T. (**2019**). "Recommendations and future directions for supervised machine learning in psychiatry," Translational Psychiatry **9**(1), 271, doi: 10.1038/s41398-019-0607-2.

Chawla, N. V., Bowyer, K. W., Hall, L. O., et Kegelmeyer, W. P. (**2002**). "SMOTE : Synthetic Minority Over-sampling Technique," Journal of Artificial Intelligence Research **16**, 321–357, doi: 10.1613/jair.953.

Cho, S. M., Austin, P. C., Ross, H. J., Abdel-Qadir, H., Chicco, D., Tomlinson, G., Taheri, C., Foroutan, F., Lawler, P. R., Billia, F., Gramolini, A., Epelman, S., Wang, B., et Lee, D. S. (**2021**). "Machine Learning Compared With Conventional Statistical Models for Predicting

Myocardial Infarction Readmission and Mortality : A Systematic Review," Canadian Journal of Cardiology **37**(8), 1207–1214, doi: 10.1016/j.cjca.2021.02.020.

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., et Van Calster, B. (**2019**). "A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models," Journal of Clinical Epidemiology **110**, 12–22, doi: 10.1016/j.jclinepi.2019.02.004.

Desai, R. J., Wang, S. V., Vaduganathan, M., Evers, T., et Schneeweiss, S. (**2020**). "Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes," JAMA Network Open **3**(1), e1918962, doi: 10.1001/jamanetworkopen.2019.18962.

Di, Y., Wang, J., Liu, X., et Zhu, T. (**2021**). "Combining Polygenic Risk Score and Voice Features to Detect Major Depressive Disorders," Frontiers in Genetics **12**, 761141, doi: 10.3389/fgene.2021.761141.

Du, Z., Li, W., Huang, D., et Wang, Y. (**2018**). "Bipolar Disorder Recognition via Multi-scale Discriminative Audio Temporal Representation," dans *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, AVEC'18, Association for Computing Machinery, New York, NY, USA, pp. 23–30, doi: 10.1145/3266302.3268997.

Faes, L., Sim, D. A., van Smeden, M., Held, U., Bossuyt, P. M., et Bachmann, L. M. (**2022**). "Artificial Intelligence and Statistics : Just the Old Wine in New Wineskins ?," Frontiers in Digital Health **4**, 833912, doi: 10.3389/fdgth.2022.833912.

Frances, A. (**2013**). "Saving normal : An insider's look at what caused the epidemic of mental illness and how to cure it," New York, NY : William Morrow .

Fried, E. I. (**2017**). "The 52 symptoms of major depression : Lack of content overlap among seven common depression scales," Journal of Affective Disorders **208**, 191–197, doi: 10.1016/j.jad.2016.10.019.

Fried, E. I., et Nesse, R. M. (**2015**). "Depression is not a consistent syndrome : An investigation of unique symptom patterns in the STAR*D study," Journal of Affective Disorders **172**, 96–102, doi: 10.1016/j.jad.2014.10.010.

Fushiki, T. (**2011**). "Estimation of prediction error by using K-fold cross-validation," Statistics and Computing **21**(2), 137–146, doi: 10.1007/s11222-009-9153-8.

Gauld, C. (**2021**). *LES RÉSEAUX DE SYMPTÔMES EN PSYCHOPATHOLOGIE Enjeux théoriques, méthodologiques et sémiologiques*.

Gauld, C., Lopez, R., Morin, C., Geoffroy, P. A., Maquet, J., Desvergnes, P., McGonigal, A., Dauvilliers, Y., Philip, P., Dumas, G., et Micoulaud-Franchi, J.-A. (**2021**). "Symptom network analysis of the sleep disorders diagnostic criteria based on the clinical text of the ICSD-3," Journal of Sleep Research **0**(0), e13435, doi: 10.1111/jsr.13435.

Gilbody, S., Richards, D., Brealey, S., et Hewitt, C. (**2007**). "Screening for Depression in Medical Settings with the Patient Health Questionnaire (PHQ) : A Diagnostic Meta-Analysis," Journal of General Internal Medicine **22**(11), 1596–1602, doi: 10.1007/s11606-007-0333-y.

Gravesteijn, B. Y., Nieboer, D., Ercole, A., Lingsma, H. F., Nelson, D., van Calster, B., Steyerberg, E. W., Åkerlund, C., Amrein, K., Andelic, N., Andreassen, L., Anke, A., Antoni, A., Audibert, G., Azouvi, P., Azzolini, M. L., Bartels, R., Barzó, P., Beauvais, R., Beer, R., Bellander, B.-M., Belli, A., Benali, H., Berardino, M., Beretta, L., Blaabjerg, M., Bragge, P., Brazinova, A., Brinck, V., Brooker, J., Brorsson, C., Buki, A., Bullinger, M., Cabeleira, M., Caccioppola, A., Calappi, E., Calvi, M. R., Cameron, P., Lozano, G. C., Carbonara, M., Chevallard, G., Chieregato, A., Citerio, G., Cnossen, M., Coburn, M., Coles, J., Cooper, D. J., Correia, M., Čović, A., Curry, N., Czeiter, E., Czosnyka, M., Dahyot-Fizelier, C., Dawes, H., De Keyser, V., Degos, V., Della Corte, F., Boogert, H. d., Depreitere, B., ?ilvesi, u., Dixit, A., Donoghue, E., Guy-Loup Dulière, J. D., Ercole, A., Esser, P., Martin Fabricius, E. E., Feigin, Kelly Foks, V. L., Frisvold, S., Furmanov, A., Gagliardo, P., Galanaud, D., Gantner, D., Gao, G., George, P., Ghuysen, A., Giga, L., Glocker, B., Golubovic, J., Gomez, P. A., Gratz, J., Gravesteijn, B., Grossi, F., Gruen, R. L., Gupta, D., Haagsma, J. A., Haitsma, I., Helbok, R., Helseth, E., Horton, L., Huijben, J., Hutchinson, P. J., Jacobs, B., Jankowski, S., Ji-yao Jiang, M. J., Jones, K., Karan, M., Kolias, A. G., Kompanje, E., Kondziella, D., Koraropoulos, E., Koskinen, L.-O., Kovács, N., Lagares, A., Lanyon, L., Laureys, S., Lecky, F., Lefering, R., Legrand, V., Lejeune, A., Levi, L., Lightfoot, R., Lingsma, H., Maas, A. I., Castaño-León, A. M., Maegele, M., Majdan, M., Manara, A., Manley, G., Martino, C., Maréchal, H., Mattern, J., McMahon, C., Melegh, B., Menon, D., Menovsky, T., Mulazzi, D., Muraleedharan, V., Murray, L., Nair, N., Negru, A., Nelson, D., Newcombe, V., Nieboer, D., Noirhomme, Q., Nyirádi, J., Olubukola, O., Oresic, M., Ortolano, F., Palotie, A., Parizel, P. M., Payen, J.-F., Perera, N., Perlbarg, V., Persona, P., Peul, W., Piippo-Karjalainen, A., Pirinen, M., Ples, H., Polinder, S., Pomposo, I., Posti, J. P., Puybasset, L., Radoi, A., Ragauskas, A., Raj, R., Rambadagalla, M., Real, R., Rhodes, J., Richardson, S., Richter, S., Ripatti, S., Rocka, S., Roe, C., Roise, O., Rosand, J., Rosenfeld, J. V., Rosenlund, C., Rosenthal, G., Rossaint, R., Rossi, S., Rueckert, D., Rusnák, M., Sahuquillo, J., Sakowitz, O., Sanchez-Porras, R., Sandor, J., Schäfer, N., Schmidt, S., Schoechl, H., Schoonman, G., Schou, R. F., Schwendenwein, E., Sewalt, C., Skandsen, T., Smielewski, P., Sorinola, A., Stamatakis, E., Stanworth, S., Kowark, A., Stevens, R., Stewart, W., Steyerberg, E. W., Stocchetti, N., Sundström, N., Synnot, A., Takala, R., Tamás, V., Tamosuitis, T., Taylor, M. S., Ao, B. T., Tenovuo, O., Theadom, A., Thomas, M., Tibboel, D., Timmers, M., Tolias, C., Trapani, T., Tudora, C. M., Vajkoczy, P., Vallance, S., Valeinis, E., Vámos, Z., Van der Steen, G., van der Naalt, J., van Dijck, J. T., van Essen, T. A., Van Hecke, W., van Heugten, C., Van Praag, D., Vyvere, T. V., Vanhaudenhuyse, A., van Wijk, R. P., Vargiolu, A., Vega, E., Velt, K., Verheyden, J., Vespa, P. M., Vik, A., Vilcinis, R., Volovici, V., von Steinbüchel, N., Voormolen, D., Vulekovic, P., Wang, K. K., Wiegers, E., Williams, G., Wilson, L., Winzeck, S., Wolf, S., Yang, Z., Ylén, P., Younsi, A., Zeiler, F. A., Zelinkova, V., Ziverte, A., et Zoerle, T. (**2020**). "Machine learning algorithms performed no better than regression models for prognostication in traumatic brain injury," Journal of Clinical Epidemiology **122**, 95–107, doi: 10.1016/j.jclinepi.2020.03.005.

Hofmann, S. G., Asnaani, A., Vonk, I. J. J., Sawyer, A. T., et Fang, A. (**2012**). "The Efficacy of Cognitive Behavioral Therapy : A Review of Meta-analyses," Cognitive Therapy and Research **36**(5), 427–440, doi: 10.1007/s10608-012-9476-1.

Kaida, K., Takahashi, M., Åkerstedt, T., Nakata, A., Otsuka, Y., Haratani, T., et Fukasawa, K. (**2006**). "Validation of the Karolinska sleepiness scale against performance and EEG variables," Clinical Neurophysiology **117**(7), 1574–1581, doi: 10.1016/j.clinph.2006.03.011.

Kirmayer, L. J. (**1989**). "Cultural variations in the response to psychiatric disorders and emo-

tional distress," Social Science & Medicine **29**(3), 327–339, doi: 10.1016/0277-9536(89)90281-5.

Kirmayer, L. J. (**1991**). "The Place of Culture in Psychiatric Nosology : Taijin Kyofusho and DSM-III-R :," The Journal of Nervous and Mental Disease **179**(1), 19–28, doi: 10.1097/00005053-199101000-00005.

Kirmayer, L. J., Lemelson, R., et Cummings, C. A. (**2015**). *Re-visioning psychiatry : Cultural phenomenology, critical neuroscience, and global mental health* (Cambridge University Press).

Kohavi, R., et others (**1995**). "A study of cross-validation and bootstrap for accuracy estimation and model selection," dans *Ijcai*, Montreal, Canada, Vol. 14, pp. 1137–1145.

Kroenke, K., Strine, T. W., Spitzer, R. L., Williams, J. B., Berry, J. T., et Mokdad, A. H. (**2009**). "The PHQ-8 as a measure of current depression in the general population," Journal of affective disorders **114**(1-3), 163–173.

Low, C. A., Dey, A. K., Ferreira, D., Kamarck, T., Sun, W., Bae, S., et Doryab, A. (**2017**). "Estimation of Symptom Severity During Chemotherapy From Passively Sensed Data : Exploratory Study," Journal of Medical Internet Research **19**(12), e420, doi: 10.2196/jmir.9046.

Lynam, A. L., Dennis, J. M., Owen, K. R., Oram, R. A., Jones, A. G., Shields, B. M., et Ferrat, L. A. (**2020**). "Logistic regression has similar performance to optimised machine learning algorithms in a clinical setting : application to the discrimination between type 1 and type 2 diabetes in young adults," Diagnostic and Prognostic Research **4**(1), 6, doi: 10.1186/s41512-020-00075-2.

Magiorkinis, E., Diamantis, A., Mitsikostas, D.-D., et Androutsos, G. (**2009**). "Headaches in antiquity and during the early scientific era," Journal of Neurology **256**(8), 1215–1220, doi: 10.1007/s00415-009-5085-7.

Marcot, B. G., et Hanea, A. M. (**2021**). "What is an optimal value of k in k-fold cross-validation in discrete Bayesian network analysis ?," Computational Statistics **36**(3), 2009–2031, doi: 10.1007/s00180-020-00999-9.

Martin, V. P., Reynal, S., Basaran, D., et Crayencour, H. C. (**2019**). "Belief Propagation algorithm for Automatic Chord Estimation," dans *16th Sound & Music Computing Conference*, pp. 537–544.

McElroy, E., Napoleone, E., Wolpert, M., et Patalay, P. (**2019**). "Structure and Connectivity of Depressive Symptom Networks Corresponding to Early Treatment Response," eClinicalMedicine **8**, 29–36, doi: 10.1016/j.eclinm.2019.02.009.

Nie, L., Zhang, L., Yang, Y., Wang, M., Hong, R., et Chua, T.-S. (**2015**). "Beyond Doctors : Future Health Prediction from Multimedia and Multimodal Observations," dans *Proceedings of the 23rd ACM international conference on Multimedia*, ACM, Brisbane Australia, pp. 591–600, doi: 10.1145/2733373.2806217.

Norman, G. R. (**2000**). "The epistemology of clinical reasoning : perspectives from philosophy, psychology, and neuroscience," Academic Medicine **75**(10), S127–S133.

Nusinovici, S., Tham, Y. C., Chak Yan, M. Y., Wei Ting, D. S., Li, J., Sabanayagam, C., Wong, T. Y., et Cheng, C.-Y. (**2020**). "Logistic regression was as good as machine learning for predicting major chronic diseases," Journal of Clinical Epidemiology **122**, 56–69, doi: 10.1016/j.jclinepi.2020.03.002.

Philip, P., Sagaspe, P., Taillard, J., Chaumet, G., Bayon, V., Coste, O., Bioulac, B., et Guilleminault, C. (**2008**). "Maintenance of Wakefulness Test, obstructive sleep apnea syndrome, and driving risk," Annals of Neurology **64**(4), 410–416, doi: 10.1002/ana.21448.

Rudin, C. (**2019**). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," Nature Machine Intelligence **1**(5), 206–215, doi: 10.1038/s42256-019-0048-x.

Rümke, H. C. (**1958**). "Die klinische Differenzierung innerhalb der Gruppe der Schizophrenien.," Der Nervenarzt .

Schultebraucks, K., Yadav, V., Shalev, A. Y., Bonanno, G. A., et Galatzer-Levy, I. R. (**2020**). "Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood," Psychological Medicine 1–11, doi: 10.1017/S0033291720002718.

Serafino, A.-M. (**2020**). "Le Praecox Feeling : présentation, revue de la littérature et aspects épistémologiques," 39.

Shinohara, S., Nakamura, M., Omiya, Y., Higuchi, M., Hagiwara, N., Mitsuyoshi, S., Toda, H., Saito, T., Tanichi, M., Yoshino, A., et Tokuno, S. (**2021**). "Depressive Mood Assessment Method Based on Emotion Level Derived from Voice : Comparison of Voice Features of Individuals with Major Depressive Disorders and Healthy Controls," International Journal of Environmental Research and Public Health **18**(10), 5435, doi: 10.3390/ijerph18105435.

Sibel, J.-C., Reynal, S., et Declercq, D. (**2012**). "A novel region graph construction based on trapping sets for the Generalized Belief Propagation," dans *2012 IEEE International Conference on Communication Systems (ICCS)*, IEEE, Singapore, Singapore, pp. 305–309, doi: 10.1109/ICCS.2012.6406159.

Soh, J., et Singh, P. (**2020**). "Machine Learning Operations," dans *Data Science Solutions on Azure* (Apress, Berkeley, CA), pp. 259–279, doi: 10.1007/978-1-4842-6405-8_8.

Spitzer, R. L., Kroenke, K., Williams, J. B., Group, P. H. Q. P. C. S., Group, P. H. Q. P. C. S., et others (**1999**). "Validation and utility of a self-report version of PRIME-MD : the PHQ primary care study," Jama **282**(18), 1737–1744.

Vanwinckelen, G., et Blockeel, H. (**2012**). "On estimating model accuracy with repeated cross-validation," dans *BeneLearn 2012 : Proceedings of the 21st Belgian-Dutch conference on machine learning*, pp. 39–44.

Westermeyer, J. (**1985**). "Psychiatric diagnosis across cultural boundaries," American Journal of Psychiatry **142**(7), 798–805, doi: 10.1176/ajp.142.7.798.

Wong, T.-T. (**2015**). "Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation," Pattern Recognition **48**(9), 2839–2846, doi: 10.1016/j.patcog.2015.03.009.

World Health Organization (**2021**). "Ethics and governance of artificial intelligence for health : WHO guidance," Rapport Technique.

Wu, P., Rallabandi, S., Black, A. W., et Nyberg, E. (**2019**). "Ordinal Triplet Loss : Investigating Sleepiness Detection from Speech," dans *Interspeech 2019*, pp. 2403–2407, doi: 10.21437/Interspeech.2019-2278.

Yedidia, J. S., Freeman, W. T., et Weiss, Y. (**2001**). "Understanding Belief Propagation and its Generalizations," .

Yedidia, J. S., Freeman, W. T., et Weiss, Y. (**2004**). "Constructing Free Energy Approximations and Generalized Belief Propagation Algorithms," .

Yigletu, H., Tucker, S., Harris, M., et Hatlevig, J. (**2004**). "Assessing suicide ideation : Comparing self-report versus clinician report," Journal of the American Psychiatric Nurses Association **10**(1), 9–15.

Zarate, C. A., et Manji, H. K. (**2009**). *Bipolar depression : molecular neurobiology, clinical diagnosis, and pharmacotherapy* (Springer).

Zimmerman, M., et McGlinchey, J. B. (**2008**). "Why don't psychiatrists use scales to measure outcome when treating depressed patients ?," The Journal of Clinical Psychiatry **69**(12), 1916–1919, doi: 10.4088/jcp.v69n1209.

Conclusion générale et perspectives

Conclusion générale

Nouvelle formulation du problème de la détection de la somnolence dans la voix

Dans ce manuscrit, nous avons proposé une nouvelle formulation du problème de la détection de la somnolence dans la voix, avec des visées d'application cliniques plutôt que d'application industrielle, comme cela était le cas pour les deux compétitions internationales de l'état de l'art. Notre travail s'est fait en collaboration étroite avec des médecins du sommeil, ce qui a imposé des contraintes d'explicabilité sur les techniques employées.

Une définition consensuelle de la somnolence

Contrairement aux systèmes existants, dont le but est de détecter la charge mentale (Boyer *et coll.*, 2016), la fatigue (Shilov et Kashevnik, 2021) ou la baisse de performances (Schuller *et coll.*, 2011, 2019) en lien avec la somnolence, nous souhaitons suivre le niveau de somnolence de patients du centre universitaire de médecine du sommeil du CHU de Bordeaux. Nous avons donc cherché à définir ce qu'est « la somnolence ». À partir de techniques de fouille de textes, d'une revue générale des outils conçus pour la mesurer et de la littérature, nous avons proposé un schéma relationnel des différents construits autour de la somnolence, en lien avec leurs mesures les plus usuelles.

Base TILE et somnolence objective

Le suivi de la somnolence de patients nécessitait la création d'une nouvelle base de données, que nous avons collectée au service universitaire de médecine du sommeil du CHU de Bordeaux. La base TILE contient ainsi les enregistrements de patients venus pour le diagnostic ou le suivi d'hypersomnies. Ils passent un Test Itératif de Latence d'Endormissement, dont la latence moyenne d'endormissement est une mesure objective de la somnolence au long cours, et remplissent de nombreux questionnaires médicaux (mesures subjectives de la somnolence). Les différences entre la base TILE et les autres corpus de l'état de l'art peuvent ainsi être identifiées selon trois axes :

— **Population** : contrairement aux corpus SLC et SLEEP qui contiennent les enregistrements de sujets de la population générale ayant au préalable rempli un questionnaire de qualité de sommeil (Buysse *et coll.*, 1989, PSQI), les sujets de la base TILE sont des patients atteints d'hypersomnie ;
— **Somnolence à court terme vs au long cours** : alors que le SLC et le SLEEP ne contiennent que des mesures de somnolence à court terme (KSS), la base TILE contient à la fois des mesures à court terme (KSS, échelle des visages) et au long cours (latence moyenne au TILE, ESS ...).
— **Objectif vs subjectif vs comportemental** : enfin, les mesures utilisées pour mesurer la somnolence dans la base TILE sont des mesures validées et utilisées en recherche et en médecine du sommeil, alors que celle utilisée pour le SLC et le SLEEP (utilisation de la KSS en auto- et hétéroquestionnaire) n'est pas validée. Par ailleurs, cette dernière mesurerait de manière dominante l'impact comportemental de la somnolence, qui comprend la voix : cette vérité terrain contient déjà – partiellement – des informations sur la voix des sujets.

Par ailleurs, une étude perceptuelle a vérifié la faisabilité de la tâche de perception de la somnolence à partir d'échantillons vocaux de ces patients, et a relié la performance des annotateurs aux niveaux d'éducation, de fatigue et d'anxiété des locuteurs.

Cette nouvelle base de données à permis une nouvelle formulation du problème de détection de la somnolence dans la voix sous la forme de « la détection de la somnolence objective (latence moyenne au TILE) et subjective (ESS, KSS moyenne) au long cours de patients hypersomniaques. »

La voix comme nouvelle mesure du TILE

Par ailleurs, l'enregistrement vocal des patients avant chaque itération du TILE peut être vu comme s'inscrivant dans les nouvelles pistes de recherches mentionnées dans le chapitre 4, consistant à trouver de nouveaux marqueurs en conservant le test standardisé du TILE tel qu'il est déjà implémenté dans les cliniques du sommeil (Pizza *et coll.*, 2019; Drakatos *et coll.*, 2013; Kawai *et coll.*, 2020; Murer *et coll.*, 2017). Cette approche propose en effet un bon compromis entre l'utilisation de tests avec lesquels les médecins du sommeil sont habitués et dans lesquels ils ont confiance, et la nouveauté proposée par l'utilisation de biomarqueurs vocaux de la somnolence. Elle pourrait ainsi être un élément permettant la sortie de la dépendance au chemin des outils utilisés en clinique, proposant une transition vers une médecine du sommeil numérique, écologique et accessible au grand public.

Symptômes, syndromes, réseaux de symptômes

Enfin, après avoir proposé dans un premier temps des classifieurs de symptômes liés à la somnolence, interprétables et en suivant une méthodologie pour leur implémentation en conditions cliniques, nous avons poussé le raisonnement plus loin et nous sommes interrogés sur l'adéquation entre la formulation des tâches d'apprentissage automatique et la réalité du raisonnement clinique.

Nous avons donc proposé dans un premier temps une réflexion sur la détection de syndromes, construits à partir des symptômes précédemment estimés ; puis nous avons proposé un ensemble de perspectives de recherche impliquant les réseaux de symptômes, permettant la formulation de nouvelles problématiques en médecine numérique.

Nouvelle dimension de marqueurs

Un deuxième axe que nous avons exploré est l'élaboration de nouveaux *biomarqueurs* pour la détection de la somnolence dans la voix.

Nouveaux biomarqueurs vocaux de la somnolence

En effet, les marqueurs les plus utilisés dans la littérature sont extraits avec la boîte à outils `openSMILE` (Eyben et Schuller, 2015). Si leur intégration dans une boîte à outils publique favorise la reproductibilité des marqueurs, ceux-ci sont très nombreux et rarement interprétables.

Nouveaux marqueurs acoustiques Nous avons donc proposé notre propre ensemble de marqueurs, qui ne contient que des mesures que nous avons validées avec des médecins du sommeil, dans le but de favoriser la future acceptation par le monde médical du système élaboré. Cet ensemble contient 47 marqueurs qui sont interprétables qu'il est possible de relier aux mécanismes de la production de parole présentés dans le chapitre 1.

Erreurs de lecture et des STA Ensuite, nous avons exploré une nouvelle dimension de l'influence de la somnolence sur la voix des patients durant leur lecture d'un texte à voix haute : les erreurs de lecture. À l'aide d'étudiants en orthophonie, nous avons annoté la base TILE et montré que certains types d'erreurs de lecture pouvaient être reliés à la somnolence des sujets. Cette nouvelle dimension, très peu étudiée dans la littérature [nous n'avons trouvé qu'une seule référence à un travail semblable pour la maladie de Parkinson (Romana *et coll.*, 2021)], est très coûteuse à la fois en temps et en expertise.

Nous avons donc, dans un deuxième temps, automatisé cette approche en reliant les erreurs faites par des systèmes de transcription automatique (STA) à la somnolence. Cette approche demande cependant des efforts supplémentaires pour l'explicabilité, les STA étant basés sur des techniques d'apprentissage profond.

Pauses de lecture : durée, nombre et emplacements Enfin, nous avons proposé une troisième dimension de la qualité de lecture des sujets, en étudiant les pauses de lecture des sujets. Alors que la durée et le nombre de pauses sont des marqueurs très utilisés dans la littérature [par ex. (Balagopalan *et coll.*, 2021; Bose *et coll.*, 2021; Clarke *et coll.*, 2021; Gonzalez-Atienza *et coll.*, 2021; Albuquerque *et coll.*, 2021; Kim *et coll.*, 2020; Demiroglu *et coll.*, 2020)], aucun travail à notre connaissance n'a étudié l'emplacement de celles-ci dans la phrase.

En combinant un système de transcription automatique et un système de détection de l'activité vocale, nous avons extrait automatiquement les durées, nombres et emplacements des pauses de lecture dans le texte lu. Ensuite, à l'aide des annotations faites sur les textes par trois étudiants en orthophonie à qui nous avons demandé de quantifier la « naturalité » des pauses entre chaque pair de mots consécutifs, nous estimons si les pauses faites par les locuteurs sont placées de manière « naturelle » ou non.

Biomarqueurs, sensibilité et spécificité

En plus de proposer de nouveaux marqueurs de la somnolence, nous avons proposé trois approches pour identifier parmi eux les *biomarqueurs* de la somnolence, c.-à-d. les marqueurs à la fois *sensibles* et *spécifiques* à la somnolence :

- Centrage par locuteur : dans les chapitres 11 et 12, nous avons proposé de centrer les marqueurs vocaux par locuteur, afin d'éliminer l'expression de leurs traits et de ne conserver que les variations d'états de ceux-ci ;
- Élimination des marqueurs corrélant avec des cofacteurs : dans le chapitre 14, nous avons proposé un système de classification dans lequel les marqueurs corrélant avec un des cofacteurs identifiés comme pouvant interférer avec la voix et/ou la somnolence (âge, sexe, IMC, tour de cou, niveau d'éducation, niveaux d'anxiété et de dépression) étaient éliminés, garantissant la spécificité des marqueurs conservés à la somnolence au regard de ces cofacteurs ;
- Décorrélation : enfin, dans le chapitre 16, nous avons proposé une première approche de décorrélation des marqueurs corrélant ces cofacteurs.

Si ces premières approches ont ouvert la voie à l'étude de la problématique des *biomarqueurs* vocaux, très peu étudiée dans la littérature, elles restent des démonstrations de faisabilité et nécessiteraient d'être approfondies, par exemple en proposant de nouvelles méthodes pour assurer la spécificité des marqueurs utilisés ; ou encore en utilisant la *triplet loss* au sein de systèmes d'apprentissages profonds.

Limites et perspectives

Modèle intégratif de l'effet de la somnolence sur la voix

Nous n'avons trouvé dans la littérature récente sur la détection de pathologies dans les voix aucun modèle intégratif de l'effet de pathologies ou de symptômes sur celle-ci, c.-à-d. un modèle explicatif à plusieurs niveaux qui lierait descripteurs vocaux, production vocale et processus neurophysiologiques. S'il existe des efforts pour relier les marqueurs vocaux aux mécanismes de production de la parole – comme proposé ici, ou par ex. dans (Hönig *et coll.*, 2014) – ces corrélations restent à des échelles comportementales, sans explication neurologique.

Un premier modèle de la production vocale a été publié par Kröger *et coll.* (2020), mais celui-ci reste limité aux processus purement moteurs et neuromoteurs, et ne prend pas en compte avec une granularité suffisamment fine les phénomènes cognitifs qui interfèrent avec celle-ci.

Un modèle intégratif de l'effet de la somnolence sur la production vocale permettrait une robustesse des descripteurs vocaux conçus pour la détecter : puisque le lien entre somnolence, mécanisme de production vocale et descripteurs vocaux est explicitée, le problème de décorrélation évoqué précédemment n'a plus lieu d'être.

Nous supposons que le développement d'un tel modèle est freiné par deux causes. D'une part, le fait que la somnolence soit elle-même mal définie (cf partie II).

D'autre part, les marqueurs extraits de la voix sont rarement reliés aux mécanismes décrits dans la partie I. S'il y a des efforts pour les marqueurs acoustiques, les marqueurs de qualité de lecture (erreurs de lecture et pauses de lecture), nécessiteraient un travail plus approfondi de neuro-psycholinguistique pour créer ce modèle intégratif. Les erreurs faites par les STA ne sont pas directement interprétables, car produites par des réseaux de neurones profonds. En revanche, un travail conjoint avec des chercheurs en intelligence artificielle explicative (*explainable AI*) permettrait peut-être d'apporter une première explication au lien entre erreurs des STA et somnolence.

D'une implémentation clinique à une implémentation écologique

Les systèmes proposés reposent sur un corpus enregistré en conditions extrêmement contrôlées (hospitalisation de jour au centre universitaire de médecine du sommeil du CHU de Bordeaux) au sein desquel les patients n'ont pas les mêmes comportements que dans leur vie courante. Une implémentation et une validation de ces systèmes en conditions écologiques semblent ainsi une nécessité.

Cependant, une des limites les plus saillantes du transfert de cette étude vers des conditions écologique est la tâche d'enregistrement proposé au sujet (lecture à voix haute). En effet, s'il est envisageable d'ajouter une tâche de lecture à haute voix d'un petit texte lors du suivi des patients par un agent conversationnel animé, la tâche risque de devenir répétitive et de faire perdre en adhérence à l'utilisation de l'application. En ce sens, il semble nécessaire d'envisager le passage vers d'autres tâches de production plus spontanées, comme la réponse à des questions ouvertes (« Comment allez-vous ? », « Comment s'est passée votre journée ? »), comme cela a déjà été observé pour d'autres pathologies (cf. chapitre 2).

Ce changement de paradigme nécessiterait une adaptation des descripteurs vocaux utilisés : si les marqueurs acoustiques restent inchangés et peuvent être directement réemployés pour ces tâches, les erreurs de lecture et des STA ne bénéficient plus d'une vérité terrain fixe et n'ont plus de raison d'être. Concernant les pauses, des travaux ont déjà utilisé leur durée

lors de la parole spontanée [par ex. dans (Albuquerque *et coll.*, 2021; Schultebraucks *et coll.*, 2020; Tan *et coll.*, 2021; Farrús *et coll.*, 2021)], mais leurs emplacements dans les phrases ne sont, encore une fois, pas étudiés à notre connaissance.

Population incluse, biais et éthique

The WEIRDEST people

Une des limites principales des résultats proposés dans ce document réside dans les caractéristiques des sujets inclus dans les bases de données étudiées.

En effet, dans (Henrich *et coll.*, 2010), Henrich a introduit le concept de population *WEIRD* [1] – occidentale, éduquée, industrialisée, riche, démocratique – et a alerté sur le fait que la majeure partie des résultats publiés en psychologie et en neurosciences avait été déterminés sur cette population, qui est pourtant particulière sous de nombreux aspects, notamment psychologiques, neurologiques et sociologiques. Dans notre étude, nous avons même amplifié ce biais, en fixant des critères d'inclusion et d'exclusion portant sur la capacité à lire à voix haute un texte en français, ne sélectionnant que les sujets ayant un niveau d'éducation relativement élevé.

Cependant, comme précisé dans le chapitre 9, nous avons fait le choix de nous concentrer sur une population précise (patients hypersomniaques venant au centre universitaire de médecine de sommeil du CHU de Bordeaux pour diagnostic ou suivi) plutôt que sur la population générale, afin d'identifier les caractéristiques vocales de l'expression de la somnolence chez ces patients-là en particulier.

Néanmoins, une limite du système que nous avons conçu est l'exclusion systématique des personnes bègues ou ayant un faible niveau de lecture (dû à leur niveau d'éducation ou à des troubles dys. par exemple) : un effort pour concevoir des marqueurs insensibles au niveau de lecture devra être mis en place pour assurer l'inclusivité de l'outil final conçu.

Symptômes

De plus, dans la visée d'une généralisation du système conçu à toute population (pas nécessairement *WEIRD*), la nécessité de travailler sur les symptômes plutôt que les diagnostics, détaillée dans le chapitre 18, prend toute sa force : puisque ceux-ci sont indépendants des cultures, des sociétés et de l'époque, travailler sur les symptômes permet d'utiliser la même vérité terrain pour toutes les cohortes, qu'elles soient *WEIRD* ou non.

Accès à un support numérique et *solutionnisme technologique*

Enfin, si l'implémentation de nouveaux outils numériques sur smartphone en population générale permettrait une amélioration de la santé publique générale des sociétés *WEIRD*, l'OMS, dans son rapport sur l'éthique et la gouvernance de l'IA en médecine (World Health Organization, 2021), met en avant la problématique éthique de l'accès aux supports numériques : il est estimé que 1.2 milliards de femmes des pays à revenus faibles à moyen (contre 873 millions d'hommes) n'ont pas accès à des supports numériques pouvant se connecter à internet, car elles n'en ont pas les moyens ou n'ont pas confiance dans ces outils. Et le genre n'est ici qu'une des dimensions affectant l'accès au numérique : géographie, culture, religion, langue et génération sont autant de facteurs discriminants. Ainsi, sans effort préalable sur l'accès au numérique de ces populations, de nouveaux outils numériques pour la santé ne

1. *Western, Educated, Industrialized, Rich, and Democratic*

permettraient pas l'accès *de tous* à un meilleur suivi ou à une meilleure prise en charge, mais uniquement pour les populations qui ont accès à des supports numériques.

De plus, dans ce même rapport, l'OMS met en garde contre un *solutionnisme technologique*, « dans lequel les technologies comme l'IA sont utilisées comme des « solutions miracles » à des problèmes sociaux, structurels, économiques et institutionnels plus profonds »[2]. L'attrait des solutions innovantes et technologiques peut conduire à une surestimation des bénéfices que celles-ci pourraient apporter et à négliger les risques et les problèmes introduits par ces technologies. À l'échelle d'un pays, cela peut se traduire par des investissements publics et des politiques de santé publique déraisonnables, au détriment d'éventuelles techniques dont l'efficacité est prouvée, mais sous-financées. La recherche de solutions technologiques – et en particulier l'IA – doit rester au service de la santé, guidée par les besoins des acteurs du monde médical, et non une fin en soi.

2. "in which technologies such as AI are used as a "magic bullet" to remove deeper social, structural, economic and institutional barriers"

Bibliographie

Albuquerque, L., Valente, A. R. S., Teixeira, A., Figueiredo, D., Sa-Couto, P., et Oliveira, C. (**2021**). "Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan," PLOS ONE **16**(4), e0248842, doi: 10.1371/journal.pone.0248842.

Balagopalan, A., Eyre, B., Robin, J., Rudzicz, F., et Novikova, J. (**2021**). "Comparing Pre-trained and Feature-Based Models for Prediction of Alzheimer's Disease Based on Speech," Frontiers in Aging Neuroscience **13**, doi: 10.3389/fnagi.2021.635945.

Bose, A., Dash, N. S., Ahmed, S., Dutta, M., Dutt, A., Nandi, R., Cheng, Y., et D Mello, T. M. (**2021**). "Connected Speech Characteristics of Bengali Speakers With Alzheimer's Disease : Evidence for Language-Specific Diagnostic Markers," Frontiers in Aging Neuroscience **13**, 707628, doi: 10.3389/fnagi.2021.707628.

Boyer, S., El-Yagoubi, R., Tiberge, M., Ruiz, R., et Daurat, A. (**2016**). "Paramètres Acoustiques de la Voix et Privation de Sommeil," dans *CFA/VISHNO*.

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., et Kupfer, D. J. (**1989**). "The Pittsburgh sleep quality index : A new instrument for psychiatric practice and research," Psychiatry Research **28**(2), 193–213, doi: 10.1016/0165-1781(89)90047-4.

Clarke, N., Barrick, T., et Garrard, P. (**2021**). "A Comparison of Connected Speech Tasks for Detecting Early Alzheimer's Disease and Mild Cognitive Impairment Using Natural Language Processing and Machine Learning," Frontiers in Computer Science **3**, doi: 10.3389/fcomp.2021.634360.

Demiroglu, C., Besirli, A., Ozkanca, Y., et Çelik, S. (**2020**). "Depression level assessment from multi-lingual conversational speech data using acoustic and text features," EURASIP Journal on Audio, Speech, and Music Processing **2020**(1), 17, doi: 10.1186/s13636-020-00182-4.

Drakatos, P., Suri, A., Higgins, S. E., Ebrahim, I. O., Muza, R. T., Kosky, C. A., Williams, A. J., et Leschziner, G. D. (**2013**). "Sleep stage sequence analysis of sleep onset REM periods in the hypersomnias," Journal of Neurology, Neurosurgery & Psychiatry **84**(2), 223–227, doi: 10.1136/jnnp-2012-303578.

Eyben, F., et Schuller, B. (**2015**). "Opensmile," ACM SIGMultimedia Records **6**, 4–13.

Farrús, M., Codina-Filbà, J., et Escudero, J. (**2021**). "Acoustic and prosodic information for home monitoring of bipolar disorder," Health Informatics Journal **27**(1), 1460458220972755, doi: 10.1177/1460458220972755.

Gonzalez-Atienza, M., Peinado, A. M., et Gonzalez-Lopez, J. A. (**2021**). "An Automatic System for Dementia Detection using Acoustic and Linguistic Features," dans *IberSPEECH 2021*, pp. 265–269, doi: 10.21437/IberSPEECH.2021-56.

Henrich, J., Heine, S. J., et Norenzayan, A. (**2010**). "The weirdest people in the world ?," Behavioral and Brain Sciences **33**(2-3), 61–83, doi: 10.1017/S0140525X0999152X.

Hönig, F., Batliner, A., Nöth, E., Schnieder, S., et Krajewski, J. (**2014**). "Acoustic-Prosodic Characteristics of Sleepy Speech - Between Performance and Interpretation," dans *Speech Prosody 2014*, pp. 864–868, doi: 10.21437/SpeechProsody.2014-162.

Kawai, R., Watanabe, A., Fujita, S., Hirose, M., Esaki, Y., Arakawa, C., Iwata, N., et Kitajima, T. (**2020**). "Utility of the sleep stage sequence preceding sleep onset REM periods for the diagnosis of narcolepsy : a study in a Japanese cohort," Sleep Medicine **68**, 9–17, doi: 10.1016/j.sleep.2019.04.008.

Kim, S., Kwon, N., O'Connell, H., Fisk, N., Ferguson, S., et Bartlett, M. (**2020**). ""How are you ?" Estimation of anxiety, sleep quality, and mood using computational voice analysis," dans *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, IEEE, Montreal, QC, Canada, pp. 5369–5373, doi: 10.1109/EMBC44109.2020.9175788.

Kröger, B. J., Stille, C. M., Blouw, P., Bekolay, T., et Stewart, T. C. (**2020**). "Hierarchical Sequencing and Feedforward and Feedback Control Mechanisms in Speech Production : A Preliminary Approach for Modeling Normal and Disordered Speech," Frontiers in Computational Neuroscience **14**(573554), doi: 10.3389/fncom.2020.573554.

Murer, T., Imbach, L. L., Hackius, M., Taddei, R. N., Werth, E., Poryazova, R., Gavrilov, Y. V., Winkler, S., Waldvogel, D., Baumann, C. R., et Valko, P. O. (**2017**). "Optimizing MSLT Specificity in Narcolepsy With Cataplexy," Sleep **40**(12), zsx173, doi: 10.1093/sleep/zsx173.

Pizza, F., Barateau, L., Jaussent, I., Vandi, S., Antelmi, E., Mignot, E., Dauvilliers, Y., Plazzi, G., et Group, f. t. M. S. (**2019**). "Validation of Multiple Sleep Latency Test for the diagnosis of pediatric narcolepsy type 1," Neurology **93**(11), e1034–e1044, doi: 10.1212/WNL.0000000000008094.

Romana, A., Bandon, J., Perez, M., Gutierrez, S., Richter, R., Roberts, A., et Provost, E. M. (**2021**). "Automatically Detecting Errors and Disfluencies in Read Speech to Predict Cognitive Impairment in People with Parkinson's Disease," dans *Interspeech 2021*, pp. 1907–1911, doi: 10.21437/Interspeech.2021-1694.

Schuller, B., Batliner, A., Bergler, C., Pokorny, F. B., Krajewski, J., Cychocz, M., Vollman, R., Roelen, S.-D., Schnieder, S., Bergelson, E., Cristia, A., Seidl, A., Warlaumont, A., Yankowitz, L., Nöth, E., Amiriparian, S., Hantke, S., et Schmitt, M. (**2019**). "The INTERSPEECH 2019 Computational Paralinguistics Challenge : Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," dans *Interspeech 2019*, doi: 10.21437/Interspeech.2019-1122.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., et Krajewski, J. (**2011**). "The INTERSPEECH 2011 Speaker State Challenge," dans *Interspeech 2011*, pp. 3201–3204, doi: 10.1.1.364.4935.

Schultebraucks, K., Yadav, V., Shalev, A. Y., Bonanno, G. A., et Galatzer-Levy, I. R. (**2020**). "Deep learning-based classification of posttraumatic stress disorder and depression following trauma utilizing visual and auditory markers of arousal and mood," Psychological Medicine 1–11, doi: 10.1017/S0033291720002718.

Shilov, N., et Kashevnik, A. (**2021**). "An Effort to Detect Vehicle Driver's Drowsy State Based on the Speed Analysis," dans *2021 29th Conference of Open Innovations Association (FRUCT)*, pp. 324–329, doi: 10.23919/FRUCT52173.2021.9435466.

Tan, E. J., Meyer, D., Neill, E., et Rossell, S. L. (**2021**). "Investigating the diagnostic utility of speech patterns in schizophrenia and their symptom associations," Schizophrenia Research **238**, 91–98, doi: 10.1016/j.schres.2021.10.003.

World Health Organization (**2021**). “Ethics and governance of artificial intelligence for health : WHO guidance,” Rapport Technique.

Annexe A

100 mots les plus représentés lors de la fouille textuelle

Mot	N_{obs}	Mot	N_{obs}	Mot	N_{obs}
sleep	42381	analysis	3647	therapy	2542
patient	24288	factor	3624	adult	2476
sleepiness	19675	conclusion	3512	assessed	2468
daytime	9632	increased	3444	however	2443
osa	9116	score	3360	problem	2416
group	8400	data	3356	baseline	2326
quality	8189	depression	3305	severe	2301
scale	7631	ahi	3299	month	2270
symptom	7246	effect	3287	woman	2252
treatment	6965	among	3229	measure	2250
associated	6504	child	3221	one	2234
disorder	6117	also	3202	osas	2215
index	5974	night	3192	without	2201
method	5320	test	3060	pressure	2196
objective	5275	duration	3013	positive	2143
apnea	5153	health	3011	trial	2116
significant	4864	outcome	3011	common	2115
time	4634	total	2978	background	2109
risk	4534	performance	2868	poor	2103
questionnaire	4520	severity	2849	individual	2060
age	4472	association	2842	population	2017
control	4423	subjective	2822	work	2013
fatigue	4271	change	2790	polysomnography	1977
clinical	4203	disturbance	2738	anxiety	1943
epworth	4194	syndrome	2706	including	1942
disease	4075	day	2698	lower	1915
year	4072	narcolepsy	2697	airway	1892
obstructive	4025	life	2670	difference	1860
insomnia	3906	high	2664	related	1857
higher	3879	prevalence	2641	improvement	1851
cpap	3777	use	2636	relationship	1845
level	3711	included	2593	activity	1836
excessive	3703	two	2591		
study	3652	cognitive	2565		

TABLEAU A.1 – 100 mots de plus de 2 lettres les plus représentés dans les 10000 premiers résultats de PubMed contenant le terme “sleepiness” dans leur titre ou leur résumé.

Annexe B

Articles de revue inclus dans notre revue générale et les mesures de la somnolence correspondantes

	OUTILS MENTIONNÉS DANS LES REVUES INCLUSES								
ARTICLE	**QUES.**	**EEG**	**PSG**	**PERF**	**BEHAV-MOTOR**	**EYE**	**AUT**	**REM**	**TOT**
(Baiardi et Mondini, 2020)	VAS-S, SSS, KSS, ESS, SWAI, THAT	QEEG	MSLT, MWT	OSLER, PVT, SART, SD, RD		EB, SEM	PPM		17
(Bodkin et Manchanda, 2011)	FSS, MFI-22, BFI, FIS, FAS, MAF, MFSI, SD		MSLT, MWT		AG				11
(Borghini *et coll.*, 2014)	NASA-TLX, SWAT	QEEG				EB, SEM	HRV, SC, RF	MEG	9
(Boulos et Murray, 2010)	SSS, ESS		MSLT, MWT	PVT					5
(Carskadon, 1993)	SSS	ERP	MSLT, MEPT	PVT	YR, PG	IE, EB	PPM		10
(Cluydts *et coll.*, 2002)	VAS-S, SSS, KSS, ESS, SWAI	ERP	MSLT, MWT	PVT, SD, RD	AG, FE, HM, SV, YR	EB	PPM		18
(Cori et al. 2019)	VAS-S, KSS	QEEG	MWT	OSLER, PVT		EB			7
(Curcio *et coll.*, 2001)	VAS-S, SSS, KSS, ESS, SWAI, RDSS, ATS	AAT, ERP	MSLT, MWT, PIS, PISS, RTSW	4CRTT, FTT, MVT, PVT, SALT, WAVT		EB, SEM, SV			23

(Dauvilliers *et coll.*, 2017)	SSS, KSS, ESS, SD	FAT, KDT, VI-GALL	MSLT, MWT, 24HPSG	PVT, SD	AG, YR, IE	EB, SEM	PPM		18
(Douglas, 2005)	ESS		MSLT, MWT	OSLER					4
(Dwarakanat et Elliott, 2019)	SSS, ESS, BSI		MSLT, MWT	OSLER, RD, SD					8
(Freedman, 2012)	SSS, KSS, ESS		MSLT, MWT	OSLER, PVT					7
(Gimbada et Rodenstein, 2009)	VAS-S, SSS, KSS, ESS		MSLT, MWT	OSLER, PVT, RD	HM	EB	PPM		12
(Hirshkowitz, 2013)	ESS, BFI, SD		MSLT, MWT	RD	AG			BM	8
(Hu et Lodewijks, 2020)	VAS-S, SSS, KSS, ESS, CFS, NASA-TLX, DSSQ, LDS, BORG	QEEG		PVT, WMVT, SD, RD, FS	FE, HM, YR, IE	EB, SV	HRV, PPM		23
(Johns, 1998)	VAS-S, SSS, KSS, ESS, SWAI	ERP	MSLT, MWT, USSWS				PPM		10
(Kaplan and Gasperetti 2020)	CF, SSS, KSS, ODSI, ESS, TODSS, RSS, SWAI, BSI, SWIFT, FACES-S, PESS, FOSQ, FACES-F, SIQ, HSI, IHSS, SD		MSLT, MWT, E-PSG	PVT, SART	AG		PPM		25
(Kendzerska *et coll.*, 2014)	SSS, KSS, ESS		MSLT, MWT						5
(Lammers *et coll.*, 2020)	ESS		MSLT, 24HPSG, 32HBR	PVT, SART	AG				7

(Liu *et coll.*, 2009)	VAS-S, SSS, KSS, Road	QEEG		SD, RD	AG	EB	HRV		10
(Manni et Tartara, 2000)	SSS, ESS	AWT	MSLT, MWT						5
(Miletin et Hanly, 2003)	ESS		MSLT						2
(Monderer *et coll.*, 2020)	SSS, KSS, ESS, UNS, SNS		MSLT, MWT	OSLER, PVT					9
(Morewitz, 1988)			MSLT						1
(Murray, 2017)	SSS, KSS, ODSI, ESS, TODSS, BSI, SWIFT, SD	ERP, QEEG	MSLT, MWT	OSLER, PVT, SART, SD	AG, PG		HRV, PPM		20
(Nami, 2012)	SSS, KSS, ESS, SD		MSLT, MWT	OSLER		EB	PPM		9
(Pertenais *et coll.*, 2019)	CF, VASS, SSS, KSS, ODSI, ESS, Hob., TODSS, RSS, SWAI, RDSS, BSI, THAT, ZOGIM-A, SWIFT, Philip, PESS, FOSQ, SIQ, UNS, SNS, NSSQ, NSS, SSI, IHSS, SKLSQ, SD	KDT, VI-GALL	MSLT, MWT, 24HPSG, 32HBR	OSLER, PVT, SART, SD	AG		PPM		39
(Plante *et coll.*, 2017)	ESS, HSI, SD		MSLT, 24HPSG, E-PSG		AG				7

(Popp *et coll.*, 2017)	ESS, FSS, FIS, NFI-MS		MSLT, MWT	PVT, SART, SD	AG		PPM		11
(Roth *et coll.*, 1982)	SSS		MSLT						2
(Huang *et coll.*, 2015)	SSS, KSS, ESS	AAT, KDT, VI-GALL	MSLT, MWT	PVT			PPM		10
(Shahid *et coll.*, 2010)	VAS-S, SSS, KSS, ESS, SWAI, FACES-S, VAS-F, FSS, FAI, MFI-20, FACES-F, TSFS, BFI, FACT, CIS, FIS, CFS, FSI, PFS		MSLT, MWT						21
(Shen *et coll.*, 2006)	VAS-S, SSS, KSS, ESS, FACES-S, VAS-F, FSS, FAI, MFI-20, FACES-F, BFI, FACT, CIS, FIS, CFS	ERP, FAT	MSLT, MWT				PPM		20
(Sparrow *et coll.*, 2019)	KSS, ESS	KDT, QEEG		PVT, SD, RD	FE, PG, YR	EB, SEM	HRV, PPM	BM	15
(Weaver, 2001)	VAS-S, SSS, KSS, ESS, SWAI, RDSS, SSSA, FOSQ		MSLT, MWT, RTSW	OSLER					12
(Wise, 2006)	SSS, ESS		MSLT, MWT						4
Nb unique values	**54**	**7**	**10**	**12**	**7**	**3**	**4**	**2**	**99**

NOM	ABB.	ARTICLE ORIGINEL	CATEGORIE
Cartoon Faces	CF	(Maldonado 2004)	QUESTIONNAIRE
Visual Analog Scale - Sleepiness	VAS-S	(Monk 1987)	QUESTIONNAIRE
Stanford Sleepiness Scale	SSS	(Hoddes 1973)	QUESTIONNAIRE
Karolinska Sleepiness Scale	KSS	(Åkerstedt 1990)	QUESTIONNAIRE
Observation and Interview Based Diurnal Sleepiness Inventory	ODSI	(Onen 2016)	QUESTIONNAIRE
Epworth Sleepiness Scale	ESS	(Johns 1991)	QUESTIONNAIRE
Hobson version of the ESS	Hob.	(Hobson 2002)	QUESTIONNAIRE
Time of Day Sleepiness Scale	TODSS	(Dolan 2009)	QUESTIONNAIRE
Resistance to Sleepiness Scale	RSS	(Violani et al 2013)	QUESTIONNAIRE
Sleep-Wake Activity Inventory	SWAI	(Rosenthal et al 1993)	QUESTIONNAIRE
Rotterdam Daytime Sleepiness Scale	RDSS	(Van Knippenberg 1995)	QUESTIONNAIRE
Barcelona Sleepiness Index	BSI	(Guaita 2015)	QUESTIONNAIRE
Toronto Hospital Alertness Test	THAT	(Shapiro 2006)	QUESTIONNAIRE
ZOGIM-A	ZOGIM-A	(Shapiro 2006)	QUESTIONNAIRE
Sleepiness-Wakefulness Inability and Fatigue Test	SWIFT	(Sangal 2012)	QUESTIONNAIRE
Fatigue, Anergy, Consciousnes, Energized and Sleepiness - Sleepiness	FACES-S	(Shapiro 2002)	QUESTIONNAIRE
Accumulated Time with Sleepiness Scale	ATS	(Gillberg 1994)	QUESTIONNAIRE
Index of Daytime Sleepiness (IDS) of the Survey Screen for Prediction of Apnea (SSSA)	SSSA	(Maislin 1995)	QUESTIONNAIRE
Likelyhood to fall asleep at the wheel	Road	(Ryener 1998)	QUESTIONNAIRE
Philip version of the ESS	Philip	(Philip 2010)	QUESTIONNAIRE
Picturial Epworth Sleepiness Scale	PESS	(Ghiassi 2010)	QUESTIONNAIRE
Functional Outcomes of Sleep Questionnaire	FOSQ	(Weaver 1997)	QUESTIONNAIRE
Visual Analog Scale - Fatigue	VAS-F	(Lee 1991)	QUESTIONNAIRE
Fatigue Severity Scale	FSS	(Krupp 1989)	QUESTIONNAIRE
Fatigue Assessment Inventory	FAI	(Schwartz 1993)	QUESTIONNAIRE
Multidimensional Fatigue Inventory	MFI-20	(Smets 1995)	QUESTIONNAIRE
Fatigue, Anergy, Consciousnes, Energized and Sleepiness - Fatigue	FACES-F	(Shapiro 2002)	QUESTIONNAIRE

Toronto Sleepiness and Fatigue Scale	TSFS	[Not published] [1]	QUESTIONNAIRE
Brief Fatigue Inventory	BFI	(Mendoza 1999)	QUESTIONNAIRE
Functional Assessment of Cancer Therapy	FACT	(Yellen 1997)	QUESTIONNAIRE
Checklist Individual Strength	CIS	(Vercoulen 1994)	QUESTIONNAIRE
Fatigue Impact Scale	FIS	(Fisk 1994)	QUESTIONNAIRE
Chalder Fatigue Scale	CFS	(Chalder 1993)	QUESTIONNAIRE
Fatigue Symptom Inventory	FSI	(Hann 1998)	QUESTIONNAIRE
Piper Fatigue Scale	PFS	(Piper 1989)	QUESTIONNAIRE
NASA Task Load Index	NASA-TLX	(Hart 1988)	QUESTIONNAIRE
Dundee Stress State Questionnaire - DSSQ		(Matthews 1999)	QUESTIONNAIRE
Neurological fatigue index	NFI-MS	(Mills 2010)	QUESTIONNAIRE
Fatigue Assessment Scale	FAS	(Michielsen 2003)	QUESTIONNAIRE
Multidimensional Assessment of Fatigue	MAF	(Belza 1995)	QUESTIONNAIRE
Multidimensional Fatigue Symptom Inventory	MFSI	(Stein 1998)	QUESTIONNAIRE
Subjective Workload Assessment Technique	SWAT	(Reid 1988)	QUESTIONNAIRE
Le Duc Scale	LDS	(Leduc 2005)	QUESTIONNAIRE
BORG	BORG	(Borg 1998)	QUESTIONNAIRE
Sleep Inertia Questionnaire	SIQ	(Kanady 2015)	QUESTIONNAIRE
Ullanlinaa Narcolepsy Scale	UNS	(Hublin 1994)	QUESTIONNAIRE
Swiss Narcolepsy Scale	SNS	(Sturzenegger 2004)	QUESTIONNAIRE
Narcolepsy Symptom Status Questionnaire	NSSQ	(Mitler 1982)	QUESTIONNAIRE
Narcolepsy Severity Scale	NSS	(Dauvilliers 2017)	QUESTIONNAIRE
Standford Sleep Inventory	SSI	(Anic-Labat 1999)	QUESTIONNAIRE
Hypersomnia Severity Index	HSI	(Kaplan 2015)	QUESTIONNAIRE
Idiopathic Hypersomnia Severity Scale	IHSS	(Dauvillier 2019)	QUESTIONNAIRE
Standford KLS Questionnaire	SKLSQ	(Arnulf 2008)	QUESTIONNAIRE
Sleep diary	SD	(Lockley 1999)	QUESTIONNAIRE
Multiple Sleep Latency Test	MSLT	(Carskadon 1979)	PSG
Maintenance of Wakefulness Test	MWT	(Mitler 1982)	PSG
Polygraphic Index of Sleepiness	PIS	(Roth 1986)	PSG
Polygraphic Score of Sleepiness	PSS	(Roth 1986)	PSG
Ultrashort sleep wake schedule	USSWS	(Lavie 1991)	PSG

1. Shen J, Chatoo K, Streiner DL, Chung SA, Huterer N, Shapiro CM. A novel questionnaire measuring sleepiness and fatigue concurrently, 2010)

The Repeated Test of Sustained Wakefulness	RTSW	(Hartse 1982)	PSG
24h continuous polysomnography	24HPSG	(Broughton 1988)	PSG
32h bedrest protocol	32HBR	(Evangelista 2018)	PSG
Microsleep events during performance test	MEPT	(Carskadon 1979)	PSG
Extended PSG	E-PSG	(Uchiyama 1995)	PSG
Four-Choice Reaction Test	4CRTT	(Wilkinson 1975)	PERF
Finger-Tapping Task	FTT	(Casagrande 1997)	PERF
The Multiple Vigilance Test	MVT	(Hirshkowistz 1993)	PERF
Oxford Sleep resistance Test	OSLER	(Benett 1997)	PERF
Psychomotor Vigilance Test	PVT	(Dinges 1985)	PERF
Simulated Assembly Line Task	SALT	(Walsh 1992)	PERF
Sustained Attention to Response Task	SART	(Robertson 1997)	PERF
Wilkinson Auditory Vigilance Task	WAVT	(Wilkinson 1969)	PERF
Working memory vigilance task	WMVT	(Matthews 2010)	PERF
Simulated Driving	SD	(Heimstra 1970)	PERF
Real driving	RD	(Brown 1967)	PERF
Flight simulator	FS	(Morris 1996)	PERF
Eye blinks frequency /duration	EB	(Stern 1984)	EYE
Slow eye movements	SEM	(Torsvall 1987)	EYE
Saccadic velocity	SV	(Galley 1989)	EYE
Alpha Attenuation Test - AAT	AAT	(Stampi 1995)	EEG
Awake Maintenance Test - AWT	AWT	(Salinski 1996)	EEG
Event-related potential	ERP	(Haider 1964)	EEG
Forced Awakening Test	FAT	(Bastuji 1999)	EEG
Karolinska Drowsiness Test	KDT	(Åkerstedt 1990)	EEG
Quantified EEG during wakefulness	QEEG	(Daniel 1967)	EEG
Vigilance Algorithm Leipzig	VIGALL	(Hegerl 2008)	EEG
Actigraphy	AG	(Newman 1988)	BEHAV/MOTOR
Facial expressions	FE	(Wierwille 1994)	BEHAV/MOTOR
Head movement	HM	(Wright 2001)	BEHAV/MOTOR
Posturography	PG	(Carskadon 1993)	BEHAV/MOTOR
Speech and voice	SV	(Whitmore 1996)	BEHAV/MOTOR
Yawning rate / mouth movements	YR	(Provine 1986)	BEHAV/MOTOR
Itching/tearing eyes	IE	(Akerstedt 1990)	BEHAV/MOTOR
Heart Rate Variability	HRV	(Egelund 1982)	AUT
Skin conductance	SC	(Davies 1965)	AUT

Pupillometry	PPM	(Lowenstein 1963)	AUT
Respiration frequency	RF	(Crawford 1961)	AUT
Biological measures	BM	(Parkes 1974)	OTHER
MEG	MEG	(Tanaka 2014)	OTHER

Bibliographie

Baiardi, S., et Mondini, S. (**2020**). "Inside the clinical evaluation of sleepiness : subjective and objective tools," Sleep and Breathing **24**(1), 369–377, doi: 10.1007/s11325-019-01866-8.

Bodkin, C., et Manchanda, S. (**2011**). "Office Evaluation of the "Tired" or "Sleepy" Patient," Seminars in Neurology **31**(01), 042–053, doi: 10.1055/s-0031-1271311.

Borghini, G., Astolfi, L., Vecchiato, G., Mattia, D., et Babiloni, F. (**2014**). "Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness," Neuroscience & Biobehavioral Reviews **44**, 58–75, doi: 10.1016/j.neubiorev.2012.10.003.

Boulos, M. I., et Murray, B. J. (**2010**). "Current evaluation and management of excessive daytime sleepiness," The Canadian Journal of Neurological Sciences. Le Journal Canadien Des Sciences Neurologiques **37**(2), 167–176, doi: 10.1017/s0317167100009896.

Carskadon, M. A. (**1993**). "Evaluation of excessive daytime sleepiness," Neurophysiologie Clinique = Clinical Neurophysiology **23**(1), 91–100, doi: 10.1016/s0987-7053(05)80287-4.

Cluydts, R., De Valck, E., Verstraeten, E., et Theys, P. (**2002**). "Daytime sleepiness and its evaluation," Sleep Medicine Reviews **6**(2), 83–96, doi: 10.1053/smrv.2002.0191.

Curcio, G., Casagrande, M., et Bertini, M. (**2001**). "Sleepiness : evaluating and quantifying methods," International Journal of Psychophysiology : Official Journal of the International Organization of Psychophysiology **41**(3), 251–263, doi: 10.1016/s0167-8760(01)00138-6.

Dauvilliers, Y., Lopez, R., et Lecendreux, M. (**2017**). "French consensus. Hypersomnolence : Evaluation and diagnosis," Revue Neurologique **173**(1-2), 19–24, doi: 10.1016/j.neurol.2016.09.017.

Douglas, N. J. (**2005**). "Assessment and management of excessive daytime sleepiness," Clinical Medicine (London, England) **5**(2), 105–108, doi: 10.7861/clinmedicine.5-2-105.

Dwarakanath, A., et Elliott, M. W. (**2019**). "Assessment of Sleepiness in Drivers," Sleep Medicine Clinics **14**(4), 441–451, doi: 10.1016/j.jsmc.2019.08.003.

Freedman, N. (**2012**). "Objective and Subjective Measurement of Excessive Sleepiness," Sleep Medicine Clinics **7**(2), 219–232, doi: 10.1016/j.jsmc.2012.03.003.

Gimbada, B. M., et Rodenstein, D. (**2009**). "[Assessment of sleepiness]," Archivos De Bronconeumologia **45**(7), 349–351, doi: 10.1016/j.arbres.2008.10.002.

Hirshkowitz, M. (**2013**). "Fatigue, Sleepiness, and Safety," Sleep Medicine Clinics **8**(2), 183–189, doi: 10.1016/j.jsmc.2013.04.001.

Hu, X., et Lodewijks, G. (**2020**). "Detecting fatigue in car drivers and aircraft pilots by using non-invasive measures : The value of differentiation of sleepiness and mental fatigue," Journal of Safety Research **72**, 173–187, doi: 10.1016/j.jsr.2019.12.015.

Huang, J., Sander, C., Jawinski, P., Ulke, C., Spada, J., Hegerl, U., et Hensch, T. (**2015**). "Test-retest reliability of brain arousal regulation as assessed with VIGALL 2.0," Neuropsychiatric Electrophysiology **1**(1), 13, doi: 10.1186/s40810-015-0013-9.

Johns, M. (**1998**). "Rethinking the assessment of sleepiness," Sleep Medicine Reviews **2**(1), 3–15, doi: 10.1016/s1087-0792(98)90050-8.

Kendzerska, T. B., Smith, P. M., Brignardello-Petersen, R., Leung, R. S., et Tomlinson, G. A. (**2014**). "Evaluation of the measurement properties of the Epworth sleepiness scale : a systematic review," Sleep Medicine Reviews **18**(4), 321–331, doi: 10.1016/j.smrv.2013.08.002.

Lammers, G. J., Bassetti, C. L., Dolenc-Groselj, L., Jennum, P. J., Kallweit, U., Khatami, R., Lecendreux, M., Manconi, M., Mayer, G., Partinen, M., Plazzi, G., Reading, P. J., Santamaria, J., Sonka, K., et Dauvilliers, Y. (**2020**). "Diagnosis of central disorders of hypersomnolence : A reappraisal by European experts," Sleep Medicine Reviews **52**, 101306, doi: 10.1016/j.smrv.2020.101306.

Liu, C. C., Hosking, S. G., et Lenné, M. G. (**2009**). "Predicting driver drowsiness using vehicle measures : Recent insights and future challenges," Journal of Safety Research **40**(4), 239–245, doi: 10.1016/j.jsr.2009.04.005.

Manni, R., et Tartara, A. (**2000**). "Evaluation of sleepiness in epilepsy," Clinical Neurophysiology : Official Journal of the International Federation of Clinical Neurophysiology **111 Suppl 2**, S111–114, doi: 10.1016/s1388-2457(00)00410-7.

Miletin, M. S., et Hanly, P. J. (**2003**). "Measurement properties of the Epworth sleepiness scale," Sleep Medicine **4**(3), 195–199, doi: 10.1016/s1389-9457(03)00031-5.

Monderer, R., Ahmed, I. M., et Thorpy, M. (**2020**). "Evaluation of the Sleepy Patient," Sleep Medicine Clinics **15**(2), 155–166, doi: 10.1016/j.jsmc.2020.02.004.

Morewitz, J. H. (**1988**). "Evaluation of excessive daytime sleepiness in the elderly," Journal of the American Geriatrics Society **36**(4), 324–330, doi: 10.1111/j.1532-5415.1988.tb02359.x.

Murray, B. J. (**2017**). "Subjective and Objective Assessment of Hypersomnolence," Sleep Medicine Clinics **12**(3), 313–322, doi: 10.1016/j.jsmc.2017.03.007.

Nami, M. T. (**2012**). "Evaluation of the sleepy patient," Australian Family Physician **41**(10), 787–790.

Pertenais, C., Lopez, R., Guichard, K., Dauvilliers, Y., Philip, P., Jaussent, I., et Micoulaud-Franchi, J.-A. (**2019**). "Revue de la littérature des outils psychométriques d'évaluation de la somnolence, de l'hypersomnolence et des hypersomnies chez l'adulte," Médecine du Sommeil **16**(4), 238–253, doi: 10.1016/j.msom.2019.08.001.

Plante, D. T., Cook, J. D., et Goldstein, M. R. (**2017**). "Objective measures of sleep duration and continuity in major depressive disorder with comorbid hypersomnolence : a primary investigation with contiguous systematic review and meta-analysis," Journal of Sleep Research **26**(3), 255–265, doi: 10.1111/jsr.12498.

Popp, R. F. J., Fierlbeck, A. K., Knüttel, H., König, N., Rupprecht, R., Weissert, R., et Wetter, T. C. (**2017**). "Daytime sleepiness versus fatigue in patients with multiple sclerosis : A systematic review on the Epworth sleepiness scale as an assessment tool," Sleep Medicine Reviews **32**, 95–108, doi: 10.1016/j.smrv.2016.03.004.

Roth, T., Roehrs, T., et Zorick, F. (**1982**). "Sleepiness : its measurement and determinants," Sleep **5 Suppl 2**, S128–134, doi: 10.1093/sleep/5.s2.s128.

Shahid, A., Shen, J., et Shapiro, C. M. (**2010**). "Measurements of sleepiness and fatigue," Journal of Psychosomatic Research **69**(1), 81–89, doi: 10.1016/j.jpsychores.2010.04.001.

Shen, J., Barbera, J., et Shapiro, C. M. (**2006**). "Distinguishing sleepiness and fatigue : focus on definition and measurement," Sleep Medicine Reviews **10**(1), 63–76, doi: 10.1016/j.smrv.2005.05.004.

Sparrow, A. R., LaJambe, C. M., et Van Dongen, H. P. (**2019**). "Drowsiness measures for commercial motor vehicle operations," Accident Analysis & Prevention **126**, 146–159.

Weaver, T. E. (**2001**). "Outcome measurement in sleep medicine practice and research. Part 1 : assessment of symptoms, subjective and objective daytime sleepiness, health-related quality of life and functional status," Sleep Medicine Reviews **5**(2), 103–128, doi: 10.1053/smrv.2001.0152.

Wise, M. S. (**2006**). "Objective measures of sleepiness and wakefulness : application to the real world ?," Journal of Clinical Neurophysiology : Official Publication of the American Electroencephalographic Society **23**(1), 39–49, doi: 10.1097/01.wnp.0000190416.62482.42.

Annexe C

Échelle de somnolence de Karolinska

	Français	**Anglais**
1	Parfaitement éveillé(e)	Extremely alert
2	Très éveillé(e)	Very alter
3	Éveillé(e)	Alert
4	Assez éveillé(e)	Rather alert
5	Ni éveillé(e) ni somnolent(e)	Neither alert nor sleepy
6	Un peu somnolent(e)	Some signs of sleepiness
7	Somnolent(e), mais sans effort pour rester éveillé(e)	Sleepy, but no effort to keep awake
8	Somnolent(e), mais avec des efforts pour rester éveillé(e)	Sleepy, but great effort to keep awake, fighting sleep
9	Très somnolent(e), luttant contre le sommeil	Extremely sleepy, can't keep awake
10	Extrêmement somnolent, ne peut rester éveillé	Extremely sleepy, can't keep awake

Tableau C.1 – Correspondance entre la version française et la version anglaise de l'échelle de somnolence de Karolinska (Karolinska Sleepiness Scale – KSS).

Annexe D

Textes utilisés dans les bases TME et TILE

1 Textes utilisés dans la base TME

1.1 Texte 1 - Des rats tenant conseil

Les Rats tenaient conseil, et ils délibéraient sur ce qu'ils avaient à faire pour se garantir de la griffe du Chat, qui avait déjà croqué plus des deux tiers de leur peuple. Comme chacun opinait à son tour, un des plus habiles se leva. – Je serais d'avis, dit-il d'un ton grave, qu'on attachât quelque grelot au cou de cette méchante bête. Elle ne pourra venir à nous sans que le grelot nous avertisse d'assez loin de son approche ; et comme en ce cas nous aurons tout le temps de fuir, vous concevez bien qu'il nous sera fort aisé de nous mettre, par ce moyen, à couvert de toute surprise de sa part. – Et toute l'assemblée applaudit aussitôt à la bonté de l'expédient. La difficulté fut de trouver un Rat qui voulût se hasarder à attacher le grelot : chacun s'en défendit ; l'un avait la patte blessée, l'autre la vue courte. – Je ne suis pas assez fort, – disait l'un. – Je ne sais pas bien comment m'y prendre –, disait l'autre. Tous alléguèrent diverses excuses, et si bonnes, qu'on se sépara sans rien conclure.

1.2 Texte 2 - Pourquoi ne faut-il surtout pas boire de l'eau de mer pour s'hydrater ?

En réalité, l'eau de mer est incapable de désaltérer la personne qui s'en abreuve à cause de sa salinité. Au contraire, celle-ci provoque une déshydratation par un effet d'osmose, facilitant la fuite de l'eau contenue dans les cellules.

Il s'agit d'une notion très connue chez les marins : en cas de naufrage, il est très fortement déconseillé de boire l'eau de mer pour tenter de s'hydrater sous peine de voir son corps se déshydrater plus rapidement. Le problème réside dans le fait que l'eau de mer contient de nombreux sels dissous.

En en buvant, l'eau que contiennent déjà les cellules de notre corps serait soumise à un effet d'osmose. En effet, elle franchirait la barrière membranaire de nos cellules afin de rejoindre le liquide le plus salé. Ce déséquilibre provoquerait malheureusement la mort des cellules par déshydratation.

Afin d'éliminer une trop grande quantité de sel ingérée, l'idéal serait de boire beaucoup d'eau douce afin de se réhydrater. Il s'agit d'un véritable cercle vicieux dont le seul élément positif est l'eau douce puisque vital pour notre corps. Dans le cas d'un naufrage, il est en revanche très possible de chercher à se réhydrater à l'aide de l'eau de pluie ou bien en mangeant si possible du poisson ou un simple fruit.

Évidemment, personne n'a encore inventé un processus portatif de désalinisation de l'eau de mer, mais cela pourrait arriver un jour. D'ailleurs, la désalinisation de l'eau de mer est un processus très coûteux pourtant massivement utilisé au Moyen-Orient, une contrée aride où l'eau douce est très rare.

1.3 Texte 3 - Des lunettes providentielles

Dans toute la forêt alentour, Madame la Taupe avait une très mauvaise réputation. Pensez donc, elle ne s'arrêtait jamais pour dire bonjour, et parfois même renversait les pauvres petites bêtes qui croisaient son chemin. Le plus astucieux des habitants de la forêt, le lapin Augustin, décida un jour de lui rendre visite et de lui demander franchement la raison de ce comportement.
« Madame la Taupe, dit Augustin, pourquoi êtes-vous donc toujours si grognon ? Vous ne saluez jamais personne, vous n'avez jamais un mot aimable. »
« Mais ce n'est pas de ma faute, répondit la taupe, je n'y vois rien et quand je m'aperçois de quelque chose, il est déjà trop tard. » C'était donc ça ! Augustin réunit alors tous les animaux qui décidèrent d'offrir une paire de lunettes à Madame la Taupe pour son anniversaire. Maintenant, tout le monde trouve qu'elle est la plus gentille et la plus polie des habitants de la forêt.

1.4 Texte 4 - La science vous donne une excellente raison de manger du chocolat !

Le chocolat, c'est pour le cerveau ! C'est en tout cas ce que confirme une étude récente menée par des chercheurs italiens. Si l'on connaissait les bienfaits du cacao sur la santé cardiovasculaire, il semblerait que nos fonctions cognitives soient également améliorées.

Si les bénéfices du chocolat et du cacao sur la santé cardiovasculaire sont avérés, leurs effets sur le cerveau sont moins connus. Il serait pourtant un véritable allié pour nos facultés cognitives. Le fait de manger du chocolat était plus précisément associé à une meilleure mémoire visuospatiale, une meilleure mémoire de travail et un meilleur raisonnement abstrait selon l'Université de L'Aquila, en Italie.

Ces bienfaits seraient à mettre au crédit des flavanols : 100 g de chocolat noir en contiennent environ 100 mg. La présence de ces petites molécules que vous retrouverez également dans les feuilles de thé et dans certains fruits et légumes permettrait de ralentir la détérioration due à la vieillesse d'une région spécifique du cerveau appelée gyrus denté. On considère communément que cette zone joue un rôle possible dans la mémoire, la préserver protégerait de facto la fonction intellectuelle.

Avant de commencer à utiliser cela comme excuse pour manger autant de chocolat que possible, n'oubliez pas que le chocolat contient également de la théobromine, un produit chimique toxique. Il vous faudrait manger en revanche au moins 80 barres d'un seul coup pour en subir les effets. Un carré de chocolat ou mieux, une fève de cacao non torréfiée par semaine suffiraient à bénéficier de ces bienfaits. C'est une piste thérapeutique à suivre pour contrecarrer le déclin cognitif et soutenir les capacités cognitives, en particulier chez les patients à risque.

2 Textes utilisés dans la base TILE

2.1 Texte 0

Les grandes personnes ne comprennent jamais rien toutes seules, et c'est fatigant, pour les enfants, de toujours et toujours leur donner des explications. J'ai donc dû choisir un autre métier et j'ai appris à piloter des avions. J'ai volé un peu partout dans le monde. Et la géographie, c'est exact, m'a beaucoup servi. Je savais reconnaître, du premier coup d'œil, la Chine de l'Arizona. C'est très utile, si l'on est égaré pendant la nuit. J'ai ainsi eu, au cours de ma vie, des tas de contacts avec des tas de gens sérieux. J'ai beaucoup vécu chez les grandes personnes.

Je les ai vues de très près. Ça n'a pas trop amélioré mon opinion. Quand j'en rencontrais une qui me paraissait un peu lucide, je faisais l'expérience sur elle de mon dessin numéro 1 que j'ai toujours conservé. Je voulais savoir si elle était vraiment compréhensive. Mais toujours elle me répondait : « C'est un chapeau. » Alors je ne lui parlais ni de serpents boas, ni de forêts vierges, ni d'étoiles. Je me mettais à sa portée. Je lui parlais de bridge, de golf, de politique et de cravates. Et la grande personne était bien contente de connaître un homme aussi raisonnable.

2.2 Texte 1

J'ai ainsi vécu seul, sans personne avec qui parler véritablement, jusqu'à une panne dans le désert du Sahara, il y a six ans. Quelque chose s'était cassé dans mon moteur. Et comme je n'avais avec moi ni mécanicien, ni passagers, je me préparai à essayer de réussir, tout seul, une réparation difficile. C'était pour moi une question de vie ou de mort. J'avais à peine de l'eau à boire pour huit jours. Le premier soir je me suis donc endormi sur le sable à mille milles de toute terre habitée. J'étais bien plus isolé qu'un naufragé sur un radeau au milieu de l'Océan. Alors vous imaginez ma surprise, au lever du jour, quand une drôle de petite voix m'a réveillé. Elle disait :
– S'il vous plaît... dessine-moi un mouton !
– Hein !
– Dessine-moi un mouton...
J'ai sauté sur mes pieds comme si j'avais été frappé par la foudre. J'ai bien frotté mes yeux. J'ai bien regardé. Et j'ai vu un petit bonhomme tout à fait extraordinaire qui me considérait gravement. Voilà le meilleur portrait que, plus tard, j'ai réussi à faire de lui. Mais mon dessin, bien sûr, est beaucoup moins ravissant que le modèle. Ce n'est pas ma faute. J'avais été découragé dans ma carrière de peintre par les grandes personnes, à l'âge de six ans, et je n'avais rien appris à dessiner, sauf les boas fermés et les boas ouverts.

2.3 Texte 2

Je regardai donc cette apparition avec des yeux tout ronds d'étonnement. N'oubliez pas que je me trouvais à mille milles de toute région habitée. Or mon petit bonhomme ne me semblait ni égaré, ni mort de fatigue, ni mort de faim, ni mort de soif, ni mort de peur. Il n'avait en rien l'apparence d'un enfant perdu au milieu du désert, à mille milles de toute région habitée. Quand je réussis enfin à parler, je lui dis :
– Mais... qu'est-ce que tu fais là ?
Et il me répéta alors, tout doucement, comme une chose très sérieuse :
– S'il vous plaît... dessine-moi un mouton...
Quand le mystère est trop impressionnant, on n'ose pas désobéir. Aussi absurde que cela me semblât à mille milles de tous les endroits habités et en danger de mort, je sortis de ma poche une feuille de papier et un stylographe. Mais je me rappelai alors que j'avais surtout étudié la géographie, l'histoire, le calcul et la grammaire et je dis au petit bonhomme (avec un peu de mauvaise humeur) que je ne savais pas dessiner. Il me répondit :
– Ça ne fait rien. Dessine-moi un mouton.
Comme je n'avais jamais dessiné un mouton je refis, pour lui, l'un des deux seuls dessins dont j'étais capable. Celui du boa fermé. Et je fus stupéfait d'entendre le petit bonhomme me répondre :
– Non ! Non ! Je ne veux pas d'un éléphant dans un boa.

2.4 Texte 3

Un boa c'est très dangereux, et un éléphant c'est très encombrant. Chez moi c'est tout petit. J'ai besoin d'un mouton. Dessine-moi un mouton. Alors j'ai dessiné. Il regarda attentivement, puis :
– Non ! Celui-là est déjà très malade. Fais-en un autre.
Je dessinai : Mon ami sourit gentiment, avec indulgence :
– Tu vois bien... ce n'est pas un mouton, c'est un bélier. Il a des cornes...
Je refis donc encore mon dessin. Mais il fut refusé, comme les précédents.
– Celui-là est trop vieux. Je veux un mouton qui vive longtemps.
Alors, faute de patience, comme j'avais hâte de commencer le démontage de mon moteur, je griffonnai ce dessin-ci. Et je lançai :
– Ça c'est la caisse. Le mouton que tu veux est dedans.
Mais je fus bien surpris de voir s'illuminer le visage de mon jeune juge :
– C'est tout à fait comme ça que je le voulais ! Crois-tu qu'il faille beaucoup d'herbe à ce mouton ?
– Pourquoi ?
– Parce que chez moi c'est tout petit...
– Ça suffira sûrement. Je t'ai donné un tout petit mouton.
Il pencha la tête vers le dessin :
– Pas si petit que ça...
Tiens ! Il s'est endormi... Et c'est ainsi que je fis la connaissance du petit prince. Voilà le meilleur portrait que, plus tard, j'ai réussi à faire de lui.

2.5 Texte 4

Il me fallut longtemps pour comprendre d'où il venait. Le petit prince, qui me posait beaucoup de questions, ne semblait jamais entendre les miennes. Ce sont des mots prononcés par hasard qui, peu à peu, m'ont tout révélé. Ainsi, quand il aperçut pour la première fois mon avion (je ne dessinerai pas mon avion, c'est un dessin beaucoup trop compliqué pour moi) il me demanda :
– Qu'est-ce que c'est que cette chose-là ?
– Ce n'est pas une chose. Ça vole. C'est un avion. C'est mon avion. Et j'étais fier de lui apprendre que je volais.
Alors il s'écria :
– Comment ! tu es tombé du ciel ?
– Oui, fis-je modestement.
– Ah ! ça c'est drôle...
Et le petit prince eut un très joli éclat de rire qui m'irrita beaucoup. Je désire que l'on prenne mes malheurs au sérieux. Puis il ajouta :
– Alors, toi aussi tu viens du ciel ! De quelle planète es-tu ?
J'entrevis aussitôt une lueur, dans le mystère de sa présence, et j'interrogeai brusquement :
– Tu viens donc d'une autre planète ?
Mais il ne me répondit pas. Il hochait la tête doucement tout en regardant mon avion :
– C'est vrai que, là-dessus, tu ne peux pas venir de bien loin...
Et il s'enfonça dans une rêverie qui dura longtemps. Puis, sortant mon mouton de sa poche, il se plongea dans la contemplation de son trésor.

2.6 Texte 5

Vous imaginez combien j'avais pu être intrigué par cette demi-confidence sur « les autres planètes ». Je m'efforçai donc d'en savoir plus long :
– D'où viens-tu, mon petit bonhomme ? Où est-ce « chez toi » ? Où veux-tu emporter mon mouton ?
Il me répondit après un silence méditatif :
– Ce qui est bien, avec la caisse que tu m'as donnée, c'est que, la nuit, ça lui servira de maison.
– Bien sûr. Et si tu es gentil, je te donnerai aussi une corde pour l'attacher pendant le jour. Et un piquet.
La proposition parut choquer le petit prince :
– L'attacher ? Quelle drôle d'idée !
– Mais si tu ne l'attaches pas, il ira n'importe où, et il se perdra…
Et mon ami eut un nouvel éclat de rire :
– Mais où veux-tu qu'il aille !
– N'importe où. Droit devant lui…
Alors le petit prince remarqua gravement :
– Ça ne fait rien, c'est tellement petit, chez moi !
Et, avec un peu de mélancolie, peut-être, il ajouta :
– Droit devant soi on ne peut pas aller bien loin…
J'avais ainsi appris une seconde chose très importante : C'est que sa planète d'origine était à peine plus grande qu'une maison ! Ça ne pouvait pas m'étonner beaucoup. Je savais bien qu'en dehors des grosses planètes comme la Terre, Jupiter, Mars, Vénus, auxquelles on a donné des noms, il y en a des centaines d'autres qui sont quelquefois si petites qu'on a beaucoup de mal à les apercevoir au télescope.

Annexe E

Description exhaustive du corpus TILE-93

Caractéristique	TILE moy. $\leq$ 8.0 min	TILE moy. $>$ 8.0 min	Tous
Caractéristiques physiques et socio-émographiques			
Sexe	F : 10 M : 11	F : 48 M : 24	F : 58 M : 35
Âge	34.1 (16.6) **	37.3 (13.6) **	36.6 (14.4)
Taille (cm)	1.7 (0.1) *	1.7 (0.1) *	1.7 (0.1)
Poids (kg)	68.2 (13.4)	68.9 (16.1)	68.7 (15.6)
IMC (kg/m^2)	23.4 (4.0)	24.1 (5.3)	23.9 (5.1)
Tour de cou (cm)	38.6 (3.7) **	37.6 (4.5) **	37.8 (4.4)
Niveau d'éducation	4.5 (1.9) ***	5.8 (2.7) ***	5.5 (2.6)
Somnolence à court- erme			
TILE (min.)	5.2 (3.5) ***	13.5 (5.7) ***	11.6 (6.3)
KSS (1-9)	4.4 (1.8)	4.4 (2.0)	4.4 (1.9)
Visage (0-6)	1.6 (0.8)	1.6 (0.9)	1.6 (0.9)
Mesures de somnolence excessive			
TILE moy. (min.)	5.2 (1.9) ***	13.5 (3.2) ***	11.6 (4.6)
ESS (0-24)	16.0 (4.9) ***	14.2 (4.6) ***	14.6 (4.7)
BSI (0-6)	2.4 (0.9)	2.3 (1.0)	2.3 (1.0)
Dimensions liées à la somnolence			
THAT (0-50)	22.2 (7.6)	23.0 (7.0)	22.8 (7.2)
Durée de sommeil estimé la veille (h)	7.0 (2.0)	7.3 (1.8)	7.2 (1.9)

Hobson (0-16)	4.8 (2.6) ***	3.9 (2.3) ***	4.1 (2.4)
ISI (0-28)	15.4 (6.3)	15.2 (5.2)	15.3 (5.5)
ASRS (0-24)	13.9 (4.3) ***	12.2 (5.2) ***	12.6 (5.0)
FOSQ-10 (10-40)	22.0 (6.6)	20.9 (6.7)	21.1 (6.7)
FSS (9-63)	47.1 (12.3)	49.3 (9.9)	48.8 (10.5)
Fatigue	0 : 3 1 : 18	0 : 5 1 : 67	0 : 8 1 : 85
Durée de sommeil estimé la veille (h)	7.0 (2.0)	7.3 (1.8)	7.2 (1.9)
Ronflements	0 : 18 1 : 3	0 : 55 1 : 17	0 : 73 1 : 20
Apnées observées	0 : 19 1 : 2	0 : 55 1 : 17	0 : 74 1 : 19
Tension	0 : 16 1 : 5	0 : 66 1 : 6	0 : 82 1 : 11
Pathologie	H Non spécifiée : 2 H.NEURO : 3 H.SAOS : 4 H.TDAH : 3 HI : 9	H Non spécifiée : 18 H.NEURO : 3 H.SAOS : 14 H.TDAH : 16 H.Tb Psy : 3 HI : 17 HI : 1	H Non spécifiée : 20 H.NEURO : 6 H.SAOS : 18 H.TDAH : 19 H.Tb Psy : 3 HI : 26 HI : 1
Anxiété et Dépression			
HAD-D (0-21)	6.4 (3.8)	6.7 (3.8)	6.7 (3.8)
HAD-A (0-21)	8.3 (3.9)	8.5 (4.3)	8.5 (4.2)
Addictions			
CAGE (0-4)	0.2 (0.5) **	0.5 (0.9) **	0.5 (0.8)
CDS5 (5-25)	7.0 (4.5)	6.8 (4.2)	6.8 (4.3)
Nb cigarettes par sem.	2.1 (5.2)	1.8 (5.2)	1.9 (5.2)
Nb alcool par sem.	0.2 (0.5) *	0.4 (0.8) *	0.3 (0.7)

Tableau E.1 – Caractéristiques des patients du corpus TILE-93

Mesure	Référence	Description
Caractéristiques physiques et socio-émographiques		
Niveau d'éducation	-	Nombre d'années d'études après le brevet des collèges
Somnolence à court- erme		
Latence d'endormissement au TILE	(Littner *et coll.*, 2005)	Temps (en min.) entre le début du test et l'endormissement du patient (0-20 min)
KSS	(Åkerstedt et Gillberg, 1990)	1 item sur la somnolence au cours des 10 dernières min. (1-9)
Échelle des visages	(Maldonado *et coll.*, 2004)	5 dessins de visages mesurant la somnolence (0-4)
Mesures de somnolence excessive		
Latence moyenne d'endormissement au TILE	(Littner *et coll.*, 2005)	Moyenne des cinq latences d'endormissement (0-20 min.)
Epworth Sleepiness Scale (ESS)	(Johns, 1991)	8 items à propos de la somnolence diurne (0-24)
Barcelona Sleepiness Index (BSI)	(Guaita *et coll.*, 2015)	2 items à propos de la somnolence (0-6)
Dimensions liées à la somnolence		
Toronto Hospital Alertness Test (THAT)	(Shahid *et coll.*, 2016)	10 items pour mesurer le niveauxd'alerte (0-50)
Hobson	(Hobson *et coll.*, 2002)	4 items à propos de la somnolence diurne excessive (0-16)
Insomnia Severity Index (ISI)	(Bastien *et coll.*, 2001)	7 items à propos de l'insomnie (0-28)
Partie A de l'ADHD Self-Report Scale (ASRS)	(Schweitzer *et coll.*, 2001)	6 items à propos du TDAH (0-24)
Functional Outcomes of Sleep Questionnaire (FOSQ-10)	(Weaver *et coll.*, 1997)	10 items sur l'impact de la somnolence diurne excessive sur le fonctionnement quotidien(10-40)
Fatigue Severity Scale (FSS)	(Krupp *et coll.*, 1989)	9 items sur la fatigue (9-63)
Durée de sommeil estimé la veille	-	Durée de sommeil estimé par le patient lui-même lors de la nuit précédant le TILE (h)
Anxiété et dépression		
Hospital Anxiety and Depression scale	(Zigmond et Snaith, 1983)	7 items sur la dépression, 7 items sur l'anxiété (0-21 chacun)
Addictions		
Cut-Down, Annoyed, Guilty, Eye-opener questionnaires (CAGE)	(Ewing, 1984)	4 items sur la consommation d'alcool
Cigarette Dependence Scale – version courte (CDS-5)	(Courvoisier et Etter, 2008)	5 items à propos de la dépendance à la cigarette

Tableau E.2 – Informations collectées dans la base TILE

Bibliographie

Åkerstedt, T., et Gillberg, M. (**1990**). "Subjective and objective sleepiness in the active individual.," Int J Neurosci **52**, 29–37, doi: 10.3109/00207459008994241.

Bastien, C. H., Vallières, A., et Morin, C. M. (**2001**). "Validation of the Insomnia Severity Index as an outcome measure for insomnia research," Sleep Medicine **2**(4), 297–307, doi: 10.1016/S1389-9457(00)00065-4.

Courvoisier, D., et Etter, J.-F. (**2008**). "Using item response theory to study the convergent and discriminant validity of three questionnaires measuring cigarette dependence.," Psychology of Addictive Behaviors **22**(3), 391–401.

Ewing, J. A. (**1984**). "Detecting Alcoholism : The CAGE Questionnaire," JAMA **252**(14), 1905.

Guaita, M., Salamero, M., Vilaseca, I., Iranzo, A., Montserrat, J. M., Gaig, C., Embid, C., Romero, M., Serradell, M., León, C., de Pablo, J., et Santamaria, J. (**2015**). "The Barcelona Sleepiness Index : A New Instrument to Assess Excessive Daytime Sleepiness in Sleep Disordered Breathing," Journal of clinical sleep medicine **11**(11), 1289–1298, doi: 10.5664/jcsm.5188.

Hobson, D. E., Lang, A. E., Martin, W. R. W., Razmy, A., Rivest, J., et Fleming, J. (**2002**). "Excessive daytime sleepiness and sudden-onset sleep in Parkinson disease : a survey by the Canadian Movement Disorders Group," JAMA **287**(4), 455–463.

Johns, M. W. (**1991**). "A New Method for Measuring Daytime Sleepiness : The Epworth Sleepiness Scale," Sleep **14**(6), 540–545, doi: https://doi.org/10.1093/sleep/14.6.540.

Krupp, L. B., LaRocca, N. G., Muir-Nash, J., et Steinberg, A. D. (**1989**). "The fatigue severity scale. Application to patients with multiple sclerosis and systemic lupus erythematosus," Archives of Neurology **46**(10), 1121–1123, doi: 10.1001/archneur.1989.00520460115022.

Littner, M. R., Kushida, C., Wise, M., Davila, D. G., Morgenthaler, T., Lee-Chiong, T., Hirshkowitz, M., Loube, D. L., Bailey, D., Berry, R. B., Kapen, S., et Kramer, M. (**2005**). "Practice Parameters for Clinical Use of the Multiple Sleep Latency Test and the Maintenance of Wakefulness Test," Sleep **28**(1), 113–121, doi: 10.1093/sleep/28.1.113.

Maldonado, C. C., Bentley, A. J., et Mitchell, D. (**2004**). "A Pictorial Sleepiness Scale Based on Cartoon Faces," Sleep **27**(3), 541–548, doi: 10.1093/sleep/27.3.541.

Schweitzer, J. B., Cummins, T. K., et Kant, C. A. (**2001**). "Attention-deficit/hyperactivity disorder," The Medical Clinics of North America **85**(3), 757–777.

Shahid, A., Chung, S., Maresky, L., Danish, A., Bingeliene, A., Shen, J., et Shapiro, C. (**2016**). "The Toronto Hospital Alertness Test scale : relationship to daytime sleepiness, fatigue, and symptoms of depression and anxiety," Nature and Science of Sleep 41, doi: 10.2147/NSS.S91928.

Weaver, T. E., Laizner, A. M., Evans, L. K., Maislin, G., Chugh, D. K., Lyon, K., Smith, P. L., Schwartz, A. R., Redline, S., Pack, A. I., et others (**1997**). "An instrument to measure functional status outcomes for disorders of excessive sleepiness," Sleep **20**(10), 835–843.

Zigmond, A. S., et Snaith, R. P. (**1983**). "The hospital anxiety and depression scale," Acta Psychiatrica Scandinavica **67**(6), 361–370, doi: 10.1111/j.1600-0447.1983.tb09716.x.

Annexe F

Algorithme ASIMPLS

Algorithme 1 : Entraînement de l'ASIMPLS

Entrées : $\mathbf{X}^{N\times M}$: Matrice de taille $N \times M$ contenant les marqueurs vocaux de la base d'entraînement
$\mathbf{y}^{N\times 1}$: matrice colonne de taille N contenant les classes S ou NS de chaque é de la base d'entraînement
l : nombre de composantes

Sorties : Matrices de projection $\mathbf{w}$, vecteurs de scores $\mathbf{T}$ et $\mathbf{U}$, matrices $\mathbf{P}$ et $\mathbf{Q}$

$\mathbf{w}$, $\mathbf{P}$, $\mathbf{Q}$, $\mathbf{T}$, $\mathbf{U} = []$
$\mathbf{E}_0^{M\times N} = \mathbf{X}$
$\mathbf{F}_0^{M\times 1} = \mathbf{y}$
$\mathbf{u}_0^{M\times 1} = \mathbf{y}$
pour $i = 1$ *à* l **faire**
  ${\mathbf{w}_i}^{N\times 1} = \mathbf{E}_{i-1}^T\mathbf{u}_{i-1}/(\mathbf{u}_{i-1}^T\mathbf{u}_{i-1})$
  $\mathbf{w}_i = \mathbf{w}_i/||\mathbf{w}_i||$ // *normalisation de* $\mathbf{w}_i$
  $\mathbf{t}_i^{M\times 1} = \mathbf{E}_{i-1}\mathbf{w}_i$
  $\mathrm{c}_i = \mathbf{F}_{i-1}^T\mathbf{t}_i/(\mathbf{t}_i^T\mathbf{t}_i)$
  $\mathbf{u}_i^{M\times 1} = c_i\mathbf{F}_{i-1}$
  $\mathrm{p}_i = \mathbf{u}_i^T\mathbf{t}_i/(\mathbf{u}_i^T\mathbf{u}_i)$
  $\mathrm{q}_i = F_{i-1}^T\mathbf{u}_i/(\mathbf{u}_i^T\mathbf{u}_i)$
  $\mathbf{E}_i = \mathbf{E}_{i-1} - \mathbf{t}_i\mathbf{t}_i^T\mathbf{E}_{i-1}/(\mathbf{t}_i^T\mathbf{t}_i)$ // *Mise à jour de* $\mathbf{E}$
  $\mathbf{F}_i = F_{i-1} - p_i\mathbf{t}_iq_i$ // *Mise à jour de* $\mathbf{F}$
  $\mathbf{w} = [\mathbf{w}, \mathbf{w}_i]; \mathbf{T} = [\mathbf{T}, \mathbf{t}_i]; \mathbf{U} = [\mathbf{U}, \mathbf{u}_i]; \mathbf{P} = [\mathbf{P}, p_i]; \mathbf{Q} = [\mathbf{Q}, q_i];$ // *Mise à jour des so*

Algorithme 2 : Recherche des paramètres $\mathbf{m}$ et $\mathbf{b}'$

Entrées : $\mathbf{y}^{N\times 1}$: matrice colonne de taille N contenant les classes S ou NS de chaque échantillon de la base de développement
l : nombre de composantes
$\mathbf{T}$: vecteur de score
$\mathbf{P}$: paramètres
$\mathbf{Q}$: matrice de charge

Sorties : Matrice de projection $\mathbf{m}$, coefficient de correction **b'**

1 $\mathbf{m} = \mathbf{P}*\mathbf{Q}$

2 trouver **b'** tel que : $\mathbf{b}'_{opt} = \underset{b'}{\operatorname{argmin}} \sum_{j=1}^{N} [(y_j - \hat{y}_j)^2] = \underset{b'}{\operatorname{argmin}} \sum_{j=1}^{N} [(y_j - \text{signe}(\sum_{i=0}^{l-1} (m_i \hat{\mathbf{t}}_i - \mathbf{b'})))^2]$

Algorithme 3 : Estimation des labels de la base de test

Entrées : $\mathbf{X}^{N\times M}$: Matrice de taille $N \times M$ contenant les marqueurs vocaux de la base de test
l : nombre de composantes
$\mathbf{w}$ et $\mathbf{m}$: matrices de projection
b' : facteur de correction

Sorties : $\hat{\mathbf{y}}$: labels estimés pour des échantillons de test

1 $\hat{\mathbf{S}}_0 = \mathbf{X}$

2 $\mathbf{T} = []$

3 **pour** *i=1 à l* **faire**

4 $\quad \hat{\mathbf{t}}_i = \hat{\mathbf{S}}_{i-1}\mathbf{w}_i$

5 $\quad \hat{\mathbf{S}}_i = \hat{\mathbf{S}}_{i-1} - \mathbf{t}_i \mathbf{t}_i^T \hat{\mathbf{S}}_{i-1}/(\mathbf{t}_i^T \mathbf{t}_i)$

6 $\quad \mathbf{T} = [\mathbf{T}, \hat{\mathbf{t}}_i]$

7 $\hat{\mathbf{y}} = \text{signe}(\sum_{i=0}^{l-1} m_i \hat{\mathbf{t}}_i - \mathbf{b}')$

Liste des acronymes

ACP	Analyse en composantes principales
AUC	*Area Under the ROC Curve* - Aire sous la courbe ROC
BPE	*Byte Pairs Encoding*
CFS	*Cartoon Faces Scale* - Échelle de somnolence des visages
ESS	*Epworth Sleepiness Scale* - Échelle de somnolence d'Epworth
KSS	*Karolinska Sleepiness Scale* - Échelle de somnolence de Karolinska
ICC	*Intraclass Correlation Coefficient* - Coefficient de corrélation intraclasse
IMC	Indice de masse corporelle
LDA	*Linear Discriminant Analysis* - Analyse discriminante linéaire
LOSOCV	*Leave One Speaker Out Cross-Validation* - Validation croisée excluant de la base d'entraînement chaque locuteur tour à tour
MAE	*Mean Absolute Error* - Erreur absolue moyenne
MSLT	*Multiple Sleep Latency Test* - cf. TILE
MWT	*Maintenance of Wakefulness Test* - cf. TME
RMSE	*Root Mean Square Error* - Erreur quadratique moyenne
SDE	Somnolence diurne excessive
SLC	*Sleepy Language Corpus*
STA	Système de transcription automatique
SVM	*Support Vector Machine* - Classifieur à vecteurs supports
TILE	Test itératif de latence d'endormissement
TME	Test de maintenance d'éveil
UAR	*Unweighted Average Recall* - Score de rappel non pondéré