Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Détection de la somnolence objective dans la voix

Vincent P. Martin[1] Jean-Luc Rouas[1] Pierre Philip[2]
(1) LaBRI - Univ. Bordeaux - Bordeaux INP - CNRS - UMR5800 - F-33400 Talence, France
(2) SANPSY - CNRS - USR 3413 - Univ. Bordeaux - CHU Pellegrin, F-33000 Bordeaux, France
vincent.martin@labri.fr, rouas@labri.fr, pierre.philip@u-bordeaux.fr

RÉSUMÉ

Le suivi des patients souffrant de maladies neuro-psychiatriques chroniques peut être amélioré grâce à la détection de la somnolence dans la voix. Cet article s'inspire des systèmes état-de-l'art en détection de la somnolence dans la voix pour le cas particulier de patients atteints de Somnolence Diurne Excessive (SDE). Pour cela, nous basons notre étude sur un nouveau corpus, le corpus TILE. Il diffère des autres corpora existants par le fait que les sujets enregistrés sont des patients souffrant de SDE et que leur niveau de somnolence est mesuré de manière subjective mais aussi objective. Le système proposé permet détecter la somnolence objective grâce à des paramètres vocaux simples et explicables à des non spécialistes.

ABSTRACT

Objective sleepiness detection through voice

The following-up of patients suffering from chronical neuro-psychiatric diseases can be improved by sleepiness detection through voice. This article draws from state-of-the-art systems to estimate sleepiness level from voice, for the specific case of patients suffering from Excessive Daytime Sleepiness (EDS). To this end, we base our study on a new corpus, the MSLT corpus. It differs from other existing corpora by the fact that recorded subjects suffer from EDS and that their sleepiness level is measured by both subjective and objective means. The proposed system allows to detect objective sleepiness with simple vocal markers that are explainable to non-specialists.

MOTS-CLÉS : Détection de la somnolence, Paramètres vocaux, Prosodie, Lecture de textes.

KEYWORDS: Sleepiness detection, Vocal markers, Prosody, Read speech.

1 Introduction

L'un des défis majeurs actuels du diagnostic et du traitement des maladies neuro-psychiatriques chroniques est la quantification des symptômes et le suivi des patients afin d'adapter leur traitement et de détecter précocement les risques de rechute. Une telle surveillance est possible en milieu écologique grâce à des dispositifs médicaux connectés (mesurant par exemple le poids, la pression sanguine ou l'activité physique) mais des informations pourtant cruciales telles que la fatigue ou la somnolence sont difficiles à mesurer par ces dispositifs. Des entretiens fréquents entre les médecins et les patients sont nécessaires, mais la quantité grandissante de patients ne permet pas un suivi régulier et personnalisé. Par ailleurs, les entretiens ne permettent pas de mesurer les variations des symptômes en réponse au traitement lorsque les patients sont à domicile. Ainsi est née l'idée de proposer aux patients un suivi à domicile en utilisant un médecin virtuel. Des études précédentes ont

Volume 1 : Journées d'Études sur la Parole, pages 406–414.
HAL : hal-02798565.

montré que l'utilisation d'un tel médecin virtuel est bien accepté par les patients (Philip *et al.*, 2020, 2017). Nous désirons compléter l'analyse des réponses aux questions posées par le médecin virtuel en y ajoutant l'analyse de paramètres vocaux. En effet, il semble désormais possible de détecter des indices dans la voix permettant d'évaluer l'état des locuteurs pour des tâches de suivi ou de diagnostic médical (Cummins *et al.*, 2018). Cette méthode présente de nombreux avantages, puisqu'elle n'est pas invasive et ne nécessite ni de capteurs spécifiques ni de processus complexe de calibration. De plus, elle peut être mise en place dans des environnements variés et permet un suivi régulier et non restrictif des patients.

Si des études précédentes ont montré qu'il est possible d'estimer la somnolence subjective dans la voix (Martin *et al.*, 2019; Schuller *et al.*, 2011; Cummins *et al.*, 2018), la plupart étaient basées sur le *Sleepy Large Corpus* (Schuller *et al.*, 2011) qui ne contient que des enregistrements de sujets sains. Plus récemment, le corpus *SLEEP* (Schuller *et al.*, 2019), élaboré pour le challenge Interspeech 2019, ouvre la voie à l'utilisation de l'apprentissage profond pour la détection de la somnolence dans la voix grâce à sa taille importante (16462 échantillons). Cependant, il est lui aussi composé d'enregistrements de sujets sains dont la somnolence est évaluée de manière subjective.

Puisque notre objectif est d'estimer la somnolence objective de patients souffrant de Somnolence Diurne Excessive (SDE), les précédents corpus ne correspondent pas à nos besoins. En effet, la somnolence est uniquement mesurée par le questionnaire médical subjectif *Karolinska Sleepiness Scale* (Åkerstedt & Gillberg, 1990) - KSS) ; les enregistrements sont effectués sur des tâches extrêmement variées allant de la production de voyelles tenues à la lecture de textes divers, en Allemand et Anglais, rendant les échantillons difficilement comparables ; les sujets enregistrés sont des sujets sains, alors que nous souhaitons suivre à domicile des patients souffrant de SDE, pour lesquels les somnolences subjectives et objectives ne corrèlent pas (Sangal, 1999). Pour résoudre ce problème nous avons donc enregistré notre propre corpus, le corpus TILE (*MSLT corpus* en anglais) (Martin *et al.*, 2020b), enregistré à la Clinique du Sommeil du Centre Hospitalier Universitaire de Bordeaux.

Nous présentons ici un système de détection de la somnolence objective de patients souffrant de SDE, grâce à des traits spécifiques de leur voix. La collaboration avec des médecins exige que les paramètres vocaux extraits des échantillons audios soient interprétables et reliés à un phénomène physiologique. En conséquence, plutôt que de mettre au point un système complexe, nous cherchons à étudier si de bonnes performances de classification peuvent être obtenues avec des paramètres vocaux simples et facilement interprétables.

Cet article est organisé comme suit. La Section 2 présente brièvement le corpus utilisé. Nous présentons dans la Section 3 les paramètres vocaux élaborés pour cette étude et dans la Section 4 notre méthodologie de classification. La Section 5 présente et discute les résultats. Enfin, une conclusion et des perspectives sont proposées dans la Section 6.

2 Description du corpus

Le corpus utilisé dans cette étude est une version augmentée du corpus TILE (Martin *et al.*, 2020b). Enregistré à la Clinique du Sommeil au Centre Hospitalier Universitaire de Bordeaux, il comprend actuellement les enregistrements de 99 patients ayant des plaintes de Somnolence Diurne Excessive (SDE). Il est basé sur le Test Itératif de Latence d'Endormissement - TILE (Littner *et al.*, 2005), durant lequel les patients font cinq siestes espacées de deux heures à partir de neuf heures du matin.

Les patients sont largement phénotypés et différentes caractéristiques physiques sont collectées pour chaque patient. Ces données sont complétées par les réponses à des questionnaires médicaux subjectifs de dépression, de fatigue et de somnolence à long terme.

Le principal avantage de ce corpus réside dans le fait qu'il associe à chaque échantillon audio à la fois une valeur de somnolence subjective (Échelle de Somnolence de Karolinska (Åkerstedt & Gillberg, 1990) - KSS) et une valeur objective mesurée par EEG (temps d'endormissement à chaque sieste, appelée "valeur de TILE" dans la suite). Les patients sont assignés à la classe Somnolent (S) ou Non-Somnolent (NS) suivant si la valeur de la moyenne des TILE sur les 5 itérations est inférieure ou supérieure à 8 minutes, limite médicale pour le diagnostic de la narcolepsie (Aldrich *et al.*, 1997) (TILE moyen $\pm$ écart-type : SL : 4,8 min $\pm$ 2,0 min ; NSL : 13,6 min $\pm$ 3,3 min). La même limite de 8 minutes est utilisée pour étiqueter les échantillons, indépendamment du label du locuteur.

Les enregistrements vocaux sont collectés durant la lecture d'un texte, qui est différent à chaque itération du test mais qui est le même pour tous les patients à session constante. Afin d'éviter une trop grande valence émotionnelle et pour avoir une grammaire et un vocabulaire simple, nous avons proposé aux sujets cinq textes, extraits du Petit Prince, d'environ 230 mots (moyenne : 229 mots, écart-type : 16,4 mots). Des statistiques concises du corpus sont présentées dans le Tableau 1. Nous redirigeons le lecteur vers (Martin *et al.*, 2020b) pour plus d'informations sur ce corpus.

Donnée	**Femmes**	**Hommes**	**Total**
Nombre de sujets	59	40	99
Nombre d'échantillons	295	200	495
Âge moyen (écart-type)	34,2 (11,6)	39,0 (17,1)	36,1 (14,3)
Niveau Social moyen (écart-type)	4.6 (2,4)	5,9 (2,6)	5,4 (2,6)
TILE moyenne (écart-type)	**11,8 (4,6)**	**10,3 (5,2)**	**11,2 (4,9)**
KSS moyen (écart-type)	4,2 (1,2)	4,6 (1,2)	4,4 (1,2)
Nombre de sujets somnolents - S	12	15	27
Nombre de sujets non somnolents - NS	47	25	72

TABLE 1 – Statistiques concises du corpus TILE (Martin *et al.*, 2020b)

3 Description des paramètres vocaux

Notre étroite collaboration avec des médecins impose que les paramètres vocaux extraits des échantillons audio soient facilement interprétables et reliés à des phénomènes physiologiques. Dans ce but, nous extrayons des paramètres avec deux granularités temporelles. D'une part, des paramètres provenant des échantillons en entier, en utilisant soit la détection automatique de segments vocaliques (Pellegrino & Andre-Obrecht, 2000) ou la détection automatique de segments voisés grâce à l'extraction de fréquence fondamentale (Sjölander, 2004). D'autre part, nous calculons des paramètres sur chaque segment voisé pour caractériser la régularité de la production d'harmoniques. Ces paramètres sont ensuite moyennés sur chaque échantillon.

3.1 Paramètres calculés sur la totalité de l'échantillon

Les statistiques sur la durée et la proportions de segments voisés et des voyelles reflètent le comportement global de la voix du locuteur.

Les paramètres extraits en utilisant ce paradigme sont :
- *duvoiced* : la durée totale des parties voisées (en s.)
- *pervoiced* : le pourcentage en durée des parties voisées
- *durvowel* : la durée totale des segments vocaliques (en s.)
- *pervowel* : le pourcentage en durée des segments vocaliques

Nous obtenons ainsi 4 paramètres vocaux sur les statistiques des segments voisés et vocaliques calculés sur l'ensemble de l'échantillon.

3.2 Paramètres calculés sur les parties voisées

Les paramètres vocaux extraits sur les parties voisées comprennent des mesures de fréquence fondamentale et de courbes d'intensité :
- F0MEAN : la moyenne de la fréquence fondamentale sur les segments voisés
- F0VAR : la variance de la fréquence fondamentale sur les segments voisés
- F0SLOPE : 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
- F0EXTEND : l'amplitude de la fréquence fondamentale sur un segment voisé

Les mêmes paramètres sont calculés sur les courbes d'intensité (NRJMEAN, NRJVAR, NRJMAX, NRJMIN, NRJEXTEND). Il en résulte 12 paramètres vocaux supplémentaires (6 sur F0, 6 sur l'intensité). Nous avons également calculé F0MEAN, F0VAR, NRJMEAN et NRJVAR sur les segments vocaliques, ajoutant ainsi 4 paramètres vocaux au groupe de paramètres.

Cet ensemble de paramètres est complété par des paramètres calculés avec le toolkit Matlab Covarep (Degottex *et al.*, 2014) que nous avons modifié pour les calculer seulement sur les segments voisés. Ces paramètres vocaux ont précédemment été utilisés pour caractériser les styles de chant (Rouas & Ioannidis, 2016) ou encore pour la classification d'attitudes sociales (Rouas *et al.*, 2019). Nous complétons ainsi notre ensemble de paramètres avec l'amplitude des harmoniques (H1,H2,H4), l'amplitude des formants (A1,A2,A3), leurs fréquences (F1,F2,F3,F4) et leurs bandes-passantes (B1,B2,B3,B4) ; 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 *Cepstral Peak Prominence* (CPP) ; les rapports harmoniques sur bruit dans différentes plages de fréquences (HNR05, HNR15, HNR25, HNR35). Tous ces paramètres sont moyennés sur chaque enregistrement, ce qui ajoute un total de 24 paramètres à notre ensemble de paramètres vocaux. Nous avons donc extrait un total de 44 paramètres.

4 Description de la méthodologie de classification

Contrairement aux précédents travaux sur la détection de la somnolence dans la voix (Martin *et al.*, 2019; Schuller *et al.*, 2019; Cummins *et al.*, 2018), le but de celui-ci n'est pas d'estimer la somnolence

instantanée du locuteur mais une somnolence à plus long terme.

Le TILE est un test reconnu médicalement pour le diagnostic de la narcolepsie (Littner *et al.*, 2005; Aldrich *et al.*, 1997). Lorsque la moyenne des valeurs de TILE des 5 siestes est inférieure à 8 minutes, le sujet est diagnostiqué comme narcoleptique. Même si la majorité de nos patients souffrent de maladies différentes de la narcolepsie (principalement d'hypersomnie idiopathique), nous choisissons de conserver la valeur limite de 8 minutes pour distinguer locuteurs Somnolents (S) et Non-Somnolents (NS), et ainsi obtenir une vérité terrain.

Pour estimer cette classe somnolence du locuteur, nous attribuons une classe (S ou NS) à chacun de ses cinq échantillons vocaux, indépendamment les uns des autres, avec la même limite de 8 minutes utilisée pour le label des locuteurs. Nous entraînons ensuite un classificateur à calculer les probabilités d'appartenance à la classe S et à la classe NS des échantillons (notés resp. p_i et $\overline{p_i}$ pour le $i$ème échantillon de chaque locuteur). En moyennant les probabilités 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 1.

La méthodologie pour calculer les probabilités d'appartenance aux classes de somnolence à partir des paramètres vocaux est similaire à celle exposée dans (Martin *et al.*, 2019).

Afin d'avoir des résultats pertinents au regard de la faible taille du corpus, nous appliquons la méthode du *Leave One Speaker Out Cross Validation* : à chaque itération de la validation croisée, un locuteur est exclu pour servir de test. Le résultat de sa classification est additionné dans une matrice de confusion globale, sur laquelle est calculée le taux de précision (*Unweighted Accuracy Recall - UAR*). Le reste des locuteurs est divisé en deux sous-corpus d'entraînements (80%) et de développement (20%), équilibrés en valeurs de TILE moyenne, en sexe et en âge, afin de pouvoir déterminer les paramètres vocaux et les paramètres du classificateur les plus pertinents. Par ailleurs, le jeu de données étant déséquilibré, la base d'entraînement est augmentée grâce à la méthode *SMOTE* implémenté dans le module Python *SciKit Learn* (Pedregosa *et al.*, 2011).

Lors de chaque itération de la validation croisée, la même procédure est appliquée :

1. Calcul pour chaque marqueur vocal de la corrélation (ρ de Spearman) entre le marqueur vocal et la valeur de l'itération de TILE sur l'ensemble *entraînement* + *développement*. Cela permet d'ordonner les paramètres vocaux du plus corrélé au moins corrélé avec la somnolence objective.

2. Sélection du nombre de features. Pour cela, nous calculons les performances du système (en *entraînement* vs *développement*) pour les 1, 2, ..., 44 paramètres vocaux, et gardons le nombre de features et les paramètres de classificateur fournissant les meilleures performances. Le classificateur utilisé est une Machine à Vecteurs Supports - SVM, dont les paramètres sont le noyau (linéaire ou gaussien) et les paramètres C et γ. Durant cette phase, les performances sont mesurées grâce au score F1 (moyenne géométrique de la précision et du rappel).

3. Le système précédent est entraîné sur le sous-corpus *entraînement+développement* et nous obtenons les probabilités d'appartenance aux classes de somnolence des 5 siestes du locuteur de *test*.

4. Les probabilités sont moyennées et seuillées pour obtenir la classe d'appartenance du locuteur. La matrice de confusion globale moyenne du système est mise à jour et nous poursuivons la validation croisée.

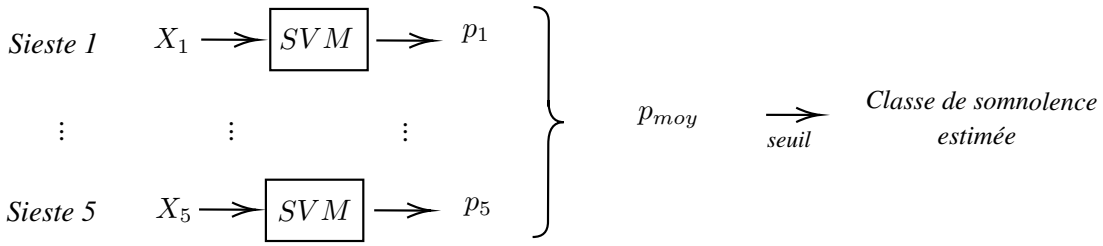

FIGURE 1 – Procédure pour l'estimation de la classe de somnolence. X_i : ensemble des paramètres vocaux triés pour la sieste i. p_i : probabilité que l'échantillon de la sieste i provienne d'un locuteur somnolent. Les classificateurs SVM représentés sont tous identiques et ont été entraînés sur tous les échantillons, toutes itérations confondues

5 Résultats et Discussions

5.1 Résultats du système

La précédente méthodologie conduit à la matrice de confusion présentée dans le Tableau 2a). Malgré les bonnes performance de cette méthode dans notre étude précédente sur la détection de la somnolence subjective (Martin *et al.*, 2019), le score UAR ainsi obtenu est seulement de 51,9%.

Pour améliorer ce système, nous décidons de modifier la procédure de sélection des paramètres vocaux : au lieu de les classer selon leur corrélation avec la valeur de TILE, nous choisissons de les classer selon leur pouvoir discriminant entre les deux classes S et NS, grâce à un test de Mann-Whitney. Plus le U de Mann-Whitney est faible, plus le marqueur vocal a des distributions différentes entre S et NS et donc un pouvoir discriminant élevé. En conservant à l'identique le reste du système, la matrice de confusion obtenue est représentée dans le Tableau 2b). Le score UAR correspondant est de 56,5%, ce qui représente une amélioration de 4,6% par rapport au système a) : cette stratégie de sélection de paramètres semble plus efficace que la précédente.

5.2 Fusion a posteriori des systèmes

Afin de tirer parti des deux différentes approches de hiérarchisation des paramètres (corrélation ou pouvoir discriminant), nous proposons une troisième stratégie. Pour chaque locuteur, nous déterminons sa probabilité d'être somnolent selon les deux systèmes, notés p^c_{moy} et p^m_{moy} (et leur équivalent pour la classe NS, $\overline{p^c_{moy}}$ et $\overline{p^m_{moy}}$). Nous effectuons la fusion des deux systèmes en prenant la classe, indépendamment du système, qui a la plus forte probabilité : $p_{moy} = max(p^c_{moy}, p^m_{moy})$ et $\overline{p_{moy}} = max(\overline{p^c_{moy}}, \overline{p^m_{moy}})$. Nous obtenons ainsi la matrice de confusion présentée dans le Tableau 2 c), conduisant à un score UAR de 60,0%, i.e. une amélioration de 3,5% par rapport au système b) et de 8,1% par rapport au système a).

5.3 Paramètres vocaux pertinents

Le meilleur système étant obtenu par la fusion de deux systèmes utilisant deux sélections de paramètres vocaux différents, il convient d'étudier quels sont les paramètres vocaux les plus pertinents pour chaque approche. Pour cela, nous séparons deux cas selon le paradigme de sélection de paramètres

a) *Corr*	S_{pred}	NS_{pred}
S_{th}	10	17
NS_{th}	24	48
UAR : 51,9%		

b) *Mann*	S_{pred}	NS_{pred}
S_{th}	11	16
NS_{th}	20	52
UAR : 56,5%		

c) *Fusion*	S_{pred}	NS_{pred}
S_{th}	11	16
NS_{th}	15	57
UAR : 60,0%		

TABLE 2 – Matrices de confusion et performances des trois systèmes étudiés : a) Tri des features par corrélation avec la valeur de somnolence (ρ de Spearman) ; b) Tri des features par leur pouvoir discriminant entre les deux classes (U de Mann-Whitney) ; c) Fusion des deux systèmes

vocaux choisi pour chaque locuteur lors de la fusion : corrélation de Spearman ou test de Mann-Whitney.

Lorsque la plus grande probabilité d'appartenance à une classe est observée pour un système de classification utilisant une corrélation de Spearman pour la sélection des features (ce qui est le cas pour 39 sujets), nous faisons la moyenne des corrélations des features avec les valeurs de TILE pour les 39 sujets concernés. Nous obtenons ainsi une corrélation moyenne des features avec les TILE pour les systèmes qui utilisent la corrélation de Spearman et qui sont efficaces dans la fusion. Les cinq paramètres les plus pertinents dans ce paradigme sont principalement des statistiques sur les parties voisées et vocaliques, ainsi que des paramètres concernant les harmoniques et les formants : CPP ($\rho = 6,7 \times 10^{-2}$), durvoiced ($\rho = 6,7 \times 10^{-2}$), durvowel ($\rho = 6,7 \times 10^{-2}$), B1 ($\rho = -6,4 \times 10^{-2}$), H1-H2 ($\rho = -6,1 \times 10^{-2}$).

Nous procédons de même avec les U de Mann-Whitney sur les 60 autres systèmes utilisant la sélection de features grâce à un test de Mann-Whitney. Les cinq paramètres vocaux les plus discriminants sont également des paramètres liés aux harmoniques et formants : H1-A1 ($U = 25981, p = 2,9 \times 10^{-2}$), H1-A2 ($U = 26186, p = 3,9 \times 10^{-2}$), CPP ($U = 26270, p = 4,3 \times 10^{-2}$), H1-H2 ($U = 26489, p = 5,9 \times 10^{-2}$), H1-A3 ($U = 26709, p = 6,1 \times 10^{-2}$).

Un point d'intérêt dans le fait de fusionner deux systèmes est l'origine des bonnes (ou mauvaises) performances de chacun des systèmes. Il est possible par exemple d'étudier l'efficacité de chaque type de système en comparant les distributions des valeurs de TILE moyenne pour les deux systèmes (les 39 sujets précédemment cités pour Spearman et 60 pour Mann-Whitney). Nous obtenons ainsi un TILE moyen pour les systèmes utilisant la discrimination par test de Mann-Whitney (moyenne : 10,3 minutes, écart-type : 5,25) significativement inférieur (test de Mann-Whitney : $U = 884,7, p = 2,1 \times 10^{-2}$) que pour les systèmes utilisant la corrélation de Spearman (moyenne : 12,4, écart-type : 4,0). Les systèmes utilisant la corrélation de Spearman sont donc plus performants sur la détection de la non-somnolence (6 Somnolents, 33 Non-Somnolents) tandis que ceux utilisant le test de Mann-Whitney comme processus de sélection de paramètres vocaux sont moins spécialisés (21 Somnolents, 39 Non-Somnolents). L'idée de faire la fusion de plusieurs systèmes est donc bénéfique dans le cadre de cette étude puisqu'elle permet de cumuler les avantages de chacun des classificateurs.

5.4 Discussion

Ce résultat est optimiste au regard de la faible taille de la base de données. Une étude plus fine de la matrice de confusion nous apporte des informations complémentaires. En effet, quel que soit le système, non seulement la majorité des sujets somnolents sont classés dans la mauvaise catégorie (60% dans le cas du meilleur système) mais la majorité des locuteurs classifiés comme somnolents

sont des patients qui ne le sont pas.

Il y a en effet un déséquilibre entre la classe NS et la classe S qui est très minoritaire. Malgré l'augmentation de données dans la classe minoritaire grâce au SMOTE, les améliorations successives du système proposé précédemment conduit à l'amélioration de la détection des patients non-somnolents plutôt que des sujets somnolents, comme en témoigne l'augmentation des performances sur cette classe là (67% pour le système (a), 72% pour le système (b) et 79,2% pour le système (c)). Un plus grand nombre de patients dans la classe S permettrait au système de mieux généraliser les caractéristiques de la voix somnolente et ainsi de surmonter ce problème.

6 Conclusion et perspectives

Pour conclure, nous avons proposé un système qui est prometteur dans la classification de la somnolence objective grâce à la voix chez des patients souffrant de Somnolence Diurne Excessive. Il s'appuie sur des paramètres vocaux qui sont explicables à des personnes qui ne sont pas spécialistes du traitement du signal vocal, et permet ainsi une collaboration étroite avec le monde médical.

Nos futurs travaux comprendront l'élaboration d'un meilleur classificateur basé sur d'autres techniques du traitement du signal, comme par exemple des techniques à base de Réseau de Neurones Récurrents pour prendre en compte les variations temporelles des paramètres vocaux. Par ailleurs, la fusion avec d'autres paradigmes de classification comme l'étude des erreurs de lecture (Martin *et al.*, 2020a) pourrait permettre de rendre le système plus robuste et d'en améliorer les performances. Enfin, la collecte de données sur des profils de patients permettant d'équilibrer les classes S et NS permettra d'obtenir des résultats plus signifiants.

Remerciements

Cette étude a été réalisée dans le cadre des projets IS-OSA, financé par la Région Nouvelle Aquitaine et SOMVOICE, financé par le Labex BRAIN (Université de Bordeaux).

Références

ALDRICH M. S., CHERVIN R. D. & MALOW B. A. (1997). Value of the multiple sleep latency test (MSLT) for the diagnosis of narcolepsy. *Sleep*, **20**(8), 620–629.

CUMMINS N., BAIRD A. & 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.

DEGOTTEX G., KANE J., DRUGMAN T., RAITIO T. & SCHERER S. (2014). COVAREP — A collaborative voice analysis repository for speech technologies. In *IEEE - ICASSP*, p. 960–964. DOI : 10.1109/ICASSP.2014.6853739.

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. & 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.

MARTIN V. P., CHAPOUTHIER G., RIEANT M., ROUAS J.-L. & PHILIP P. (2020a). Using reading mistakes as features for sleepiness detection in speech. In *The 10th International Conference on Speech Prosody*.

MARTIN V. P., ROUAS J.-L., MICOULAUD-FRANCHI J.-A. & PHILIP P. (2020b). The Objective and Subjective Sleepiness Voice Corpora. In *12th Language Resources and Evaluation Conference*.

MARTIN V. P., ROUAS J.-L., THIVEL P. & KRAJEWSKI J. (2019). Sleepiness detection on read speech using simple features. In *10th Conference on Speech Technology and Human-Computer Dialogue*. DOI : 10.1109/SPED.2019.8906577.

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. & DUCHESNAY E. (2011). Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

PELLEGRINO F. & ANDRE-OBRECHT R. (2000). Automatic language identification : an alternative approach to phonetic modelling. *Signal Processing*, **80**(7), 1231–1244.

PHILIP P., DUPUY L., AURIACOMBE M., SERRE F., DE SEVIN E., SAUTERAUD A. & MICOULAUD-FRANCHI J.-A. (2020). Trust and acceptance of a virtual psychiatric interview between embodied conversational agents and outpatients. *npj Digital Medicine*, **3**(1), 2. DOI : 10.1038/s41746-019-0213-y.

PHILIP P., MICOULAUD-FRANCHI J.-A., SAGASPE P., DE SEVIN E., OLIVE J., BIOULAC S. & 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.

ROUAS J.-L. & IOANNIDIS L. (2016). Automatic Classification of Phonation Modes in Singing Voice : Towards Singing Style Characterisation and Application to Ethnomusicological Recordings. In *Interspeech 2016*, p. 150–154.

ROUAS J.-L., SHOCHI T., GUERRY M. & RILLIARD A. (2019). Categorisation of spoken social affects in Japanese : human vs. machine. In *ICPhS*.

SANGAL R. (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.

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. & SCHMITT M. (2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge : Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. In *Interspeech 2019*.

SCHULLER B., STEIDL S., BATLINER A., SCHIEL F. & KRAJEWSKI J. (2011). The INTERSPEECH 2011 Speaker State Challenge. In *Interspeech 2011*, p. 3201–3204.

SJÖLANDER K. (2004). The Snack Sound Toolkit.

ÅKERSTEDT T. & GILLBERG M. (1990). Subjective and objective sleepiness in the active individual. *Int J Neurosci*, **52**, 29–37.