AUTOMATIC DETECTION OF SLEEPINESS-RELATED SYNDROMES AND SYMPTOMS USING VOICE AND SPEECH BIOMARKERS

Vincent P. Martin[1,2,3], *Jean-Luc Rouas*[2], *Pierre Philip*[3]

[1] DDP Research Unit, Department of Precision Health, LIH, Strassen, Luxembourg
[2] Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400 Talence, France
[3] Univ. Bordeaux, CNRS, SANPSY, UMR 6033, CHU Pellegrin, F-33076 Bordeaux, France

ABSTRACT

This article is about the automatic estimation of sleepiness in hypersomnia patients recorded during a reading task. Based on the Multiple Sleep Latency Corpus, our main contribution is to explore new formulations of sleepiness detection in speech by specifying and performing five sleepiness-related classification tasks. We automatically classify three symptoms, and two syndromes, i.e. combinations of symptoms that are closer to clinical reasoning. Another contribution of this paper is the use of a simple and interpretable pipeline integrating selecting voice biomarkers of sleepiness, i.e. features that are both sensible and specific to sleepiness. In particular, specificity is adressed integrating a decorrelation step in the pipeline, which allows to certify that the descriptors selected by the pipeline are indeed specific of sleepiness with respect to 7 cofactors (age, BMI, etc.).

Index Terms— Voice biomarkers, Excessive sleepiness, Clinical practice, Symptoms

1. INTRODUCTION

Context and state of the art. Sleepiness is a complex psychophysiological phenomenon with a high prevalence (up to 40% of the general population [1]), which has negative consequences on both personal and public health, increasing the risk of disability and mortality [2]. A promising tool to measure sleepiness in ecological conditions and passive set-ups – and thus enhance the follow-up of sleepiness by clinicians – is speech analysis.

In this way, instantaneous sleepiness detection in speech has been the focus of two international challenges, proposed in parallel to the 2011 and 2019 Interspeech conferences [3, 4]. During both challenges, the corpora have been annotated with the average of three Karolinska Sleepiness Scale (KSS), one being annotated by the subjects and the two others by external annotators. In 2011, the system achieving the best binary classification performances (Unweighted Average Recall – UAR) on the whole corpus has reached 71.7% [5] while other efforts have reached 76.4% on a subcorpus of reading tasks [6]. The winner of the 2019 challenge reached a Pearson correlation coefficient of $\rho = .387$ between estimations and ground truth of the challenge corpus [7]. This performance is still state-of-the-art despite recent works on this corpus [8, 9]. More recently, during ICASSP 2022, a new large-scale dataset has been introduced: the Voiceome dataset, containing the recordings of 1828 subjects from the general population and annotated with the Standford Sleepiness Scale (SSS).The proposed system, based on masking and separated training, reached an accuracy of 81.13% [10].

However, neither of the two scales used to annotate these datasets (resp. average of three KSS or SSS) is used in sleep medicine [11]. This hinders the use and endorsement of these tools by clinicians, who discern numerous psychophysiological constructs around sleepiness [12]. Moreover, "the clinician attempting to make a diagnosis is dealing almost exclusively at the syndrome level" [13], a *syndrome* being defined as "a cluster of signs and symptoms" [13] of higher conceptual level. To our knowledge, previous studies focused on a unique symptom of sleepiness (i.e. instantaneous subjective sleepiness), whereas during clinical interviews, clinicians investigate several symptoms and syndromes related to sleepiness [12].

Objectives. While some of our previous works have initiated a multidimensional study of long-term sleepiness in sleep clinic patients (e.g. [14, 15, 16]), our objective is to go further and propose the classification of both symptoms and syndromes related to excessive sleepiness using voice recordings.

This article is structured as follows. In Section 2, we introduce the corpus used in this study as well as the three symptoms and two syndromes that we aim to classify. We introduce our vocal features and classification pipeline in Section 3. We provide the results of classification and discuss them in Section 4. Finally, we draw conclusions in Section 5.

This research is funded by the French Research Agency ANR through the axis "Autonom-Health" of the PEPR "Santé Numérique", Grant agreement n°ANR-22-PESN-000X

2. CORPUS AND SYNDROME DEFINITIONS

Corpus. The Multiple Sleep Latency Test corpus (MSLTc) contains the recordings of 106 patients admitted to the sleep medicine department of the Bordeaux University Hospital (France) for the diagnosis and/or treatment of rare hypersomnia diseases [17, 18]. They undertake a Multiple Sleep Latency Test (MSLT), consisting of asking them to take five 20-min naps every two hours, from 9 a.m. to 5 p.m. During these naps, the time they fall asleep is assessed by a specialized nurse watching online polysomnographic recordings: this value, named *sleep latency*, is an objective measurement of short-term propensity to fall asleep in sleep-favorable conditions. The Mean Sleep Latency (MSL) across the five naps is a reference criterion in sleep medicine, measuring the average sleep propensity of a patient over a long period of time [19].

Before each nap, the patients are asked to read out loud an extract from *Le Petit Prince* (A. de Saint-Exupéry) and they fill out an instantaneous sleepiness questionnaire – the Karolinska Sleepiness Scale [20, KSS]. During their stay at the hospital, the patients are asked to fill out numerous sleep-related and health questionnaires [17], including the Epworth Sleepiness Scale [21, ESS], a questionnaire measuring the subjective perception of sleep propensity during the two previous weeks.

Symptoms and syndromes. All the recorded subjects are patients of the sleep clinic of Bordeaux (France) complaining about hypersomnolence with different etiologies and different symptomatic profiles (i.e. different combinations of symptoms). We focus here on the detection of symptoms and syndromes independently from their underlying disease. Of all the measures included in the MSLTc, we focus on the three following symptoms:

(S1) *Objective severe daytime sleep propensity*, measured by an MSL ≤ 8 min (28/106 patients, 26.4%).

(S2) *Subjective perception of severe daytime sleep propensity*, measured by the score to an ESS > 15, (45/106 patients, 42.5%).

(S3) *Average subjective perception of instantaneous sleepiness across the day*, measured by an average KSS during the MSLT > 5 (31/106 patients, 29.2%).

The number of patients belonging to each symptomatic profile is represented in Figure 1. From all these symptomatic profiles, two syndromes are of particular interest to sleep clinicians [12]:

(SP) *Sleep Propensity*, which measures the pathological tendency of patients to fall asleep when the complaint of excessive sleep propensity (ESS > 15) is objectified by an MSL ≤ 8 minutes (S1 $\cap$ S2). In this binary classification task, 16/106 patients (15.0%) are affected by the SP syndrome.

(SS) *Subjective Sleepiness*, which measures a general complaint of the subjects about excessive sleepiness. It is defined by an average KSS higher than 5 and a score to the ESS higher than 15 (S2 $\cap$ S3). In this binary classification task, 14/106 patients (13.2%) are affected by the SS syndrome.

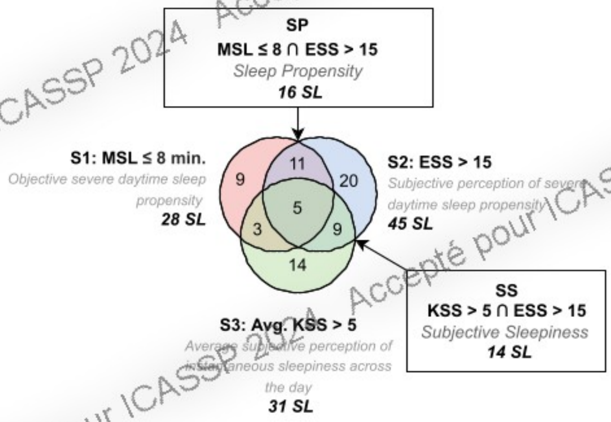
SP
MSL ≤ 8 ∩ ESS > 15
Sleep Propensity
16 SL
S1: MSL ≤ 8 min.
Objective severe daytime sleep propensity
28 SL
S2: ESS > 15
Subjective perception of severe daytime sleep propensity
45 SL
9
11
20
5
3
9
14
SS
KSS > 5 ∩ ESS > 15
Subjective Sleepiness
14 SL
S3: Avg. KSS > 5
Average subjective perception of instantaneous sleepiness across the day
31 SL

Fig. 1. Venn diagram of the symptomatic profiles of the patients in the MSLTc

3. FEATURES AND CLASSIFICATION PIPELINE

Constraints. As this work is carried out in close collaboration with sleep physicians, the interpretability of our results by them is a major criterion of our work [22]. We therefore decided to put this factor as a priority in the design of our descriptors and our classification pipeline [23]. We thus used descriptors that are understandable by non-specialists in signal processing, and tools in the classification pipeline with which they are already familiar (i.e. PCA or regression). In the same way, we did not use any deep learning-related technique for classification.

Features. The voice and speech features extracted in this study are divided into three categories:

(i) Acoustic parameters. First, we extracted acoustic features ($n = 30$) [6] validated with clinicians. They are computed on voiced parts using the `Snack` toolbox: fundamental frequency and energy average and variance, harmonics and formants amplitude and bandwidth, Harmonic to Noise ratio, ...

(ii) Automatic Speech Recognition errors. Second, similarly to previous works [14, 15], we computed Automatic Speech Recognition (ASR) errors ($n = 8$). We consider in this study four word metrics (Insertions, Substitutions, Deletions, and the number of correctly identified words – Correct) made by a conformer-based ASR system, implemented in espnet [24] and trained on ESTER [25]. The metrics are computed on words, and we consider both the raw values and the ratio over the total number of identified words.

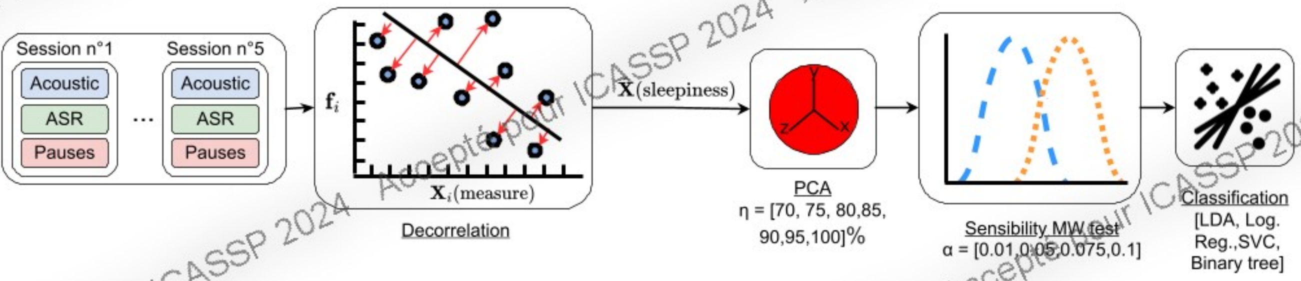

Fig. 2. Classification pipeline and hyperparameters. f_i: confounding factors, MW: Mann-Whitney, PCA: Principal Component Analysis, LDA: Linear Discriminant Analysis, Log. Reg.: Logistic Regression, SVC: Support Vector Classifier

(iii) Pauses. Finally, we extracted reading pauses duration and location ($n = 24$) [16]. Combining an ASR system and a Voice Activity Detector, we estimate the duration and location in the original text of the reading pauses. Then, depending on their location (correctly or incorrectly placed), we compute their number, duration, 'naturalness' scores (cf. [26]), and duration-weighted naturalness scores.

Features aggregation. Our aim is to detect long-term sleepiness-related symptoms and syndromes from the five recordings per patient collected during an MSLT. To do so, we test four fusion strategies: averaging them across the naps, keeping only the average and the standard deviation, keeping only the standard deviation, and aggregating all features.

Classification pipeline. Features are then fed into the pipeline represented in Figure 2. Since we collaborate with clinicians, our features and pipeline have to be explainable by design [23] so that clinicians can understand and have confidence in the proposed system [22]. Moreover, we aim at designing *biomarkers* of symptoms and syndromes related to sleepiness, i.e. voice and speech features that are both *sensible* and *specific* to sleepiness. Thus, in the same vein of previous works [14, 15], we include in the features selection pipeline a decorrelation step, in order to ensure that selected features are independent from confounding factors. Indeed, multiple traits are expressed through one's voice, and we need to ensure that the selected features are *specific* to sleepiness. In the MSLTc, seven confounding factors have been identified as potentially interfering with both sleepiness and voice: age, sex, Body Mass Index (BMI), neck circumference, anxiety and depression levels (as measured by the Hospital Anxiety and Depression scale [27]), and educational level (as measured by the highest level obtained after the French Certificate of Education), all considered as known by the clinicians. The decorrelation process is thus done outside any cross-validation loop.

The decorrelation procedure is the following. For each feature $\mathbf{X}_i$(measured), we compute a Pearson correlation test with all confounding factors, and identify the ones that correlate significantly with $\mathbf{X}_i$(measured). We denote them by $\mathbf{f}_i$. Then, we estimate the contribution of sleepiness to the value of $\mathbf{X}_i$(measured), denoted by $\hat{\mathbf{X}}_i$(sleepiness), by substracting the value of the contribution of the confounding factors (denoted by $\hat{\mathbf{X}}_i(\mathbf{f}_i)$) to $\mathbf{X}_i$(measured):

$$\hat{\mathbf{X}}_i(\text{sleepiness}) = \mathbf{X}_i(\text{measured}) - \hat{\mathbf{X}}_i(\mathbf{f}_i)$$

In a first version, we estimate $\hat{\mathbf{X}}_i(\mathbf{f}_i)$ by a multivariate linear regression between $\mathbf{X}_i$(measured) and $\mathbf{f}_i$. The features still correlating with at least one confounding factor after this procedure are excluded from the selected features since they are not specific to sleepiness.

Cross-Validation. The validation of the system is performed under nested cross-validation. The external loop is performed under Leave One Speaker Out cross-validation (LOSO), while the internal loop is evaluated under stratified k-fold with different parameters $k_{inner} \in \{3, 5, 10\}$. For each loop, predictions are aggregated before computing the Unweighted Average Recall (UAR – as recommended in guidelines [28]) on all the aggregated predictions.

While LOSO has been described as inflating performances on regression tasks [29], we used it in the external loop to have the same training and test bases for each of the symptoms and syndromes, in order to provide a valid comparison. Doing so, each testing sample corresponds to the features of a unique patient, mimicking the clinical reasoning, i.e. evaluating one patient based on the knowledge accumulated on the previous ones [30].

4. RESULTS AND DISCUSSION

Classification pipeline. The implemented classification pipeline provided satisfactory classification performances on symptoms (S1: UAR=81.5%, S2: UAR=76.6%, S3: UAR=79.0%), but more mixed performances on syndromes: while reasonable performances were obtained on the detection of SP (UAR=75.3%), SS was detected only with an UAR of 66.2%. The effect of feature aggregation is reported in Table 2 and the effect of all the other pipeline parameters are reported in Table 1.

(i) Features fusion and combination. For symptoms estimation, the best performances are obtained by averaging the features of the five recordings per speaker. On the contrary,

Ref.	k_{inner}			**Decorrelation**			**Sen.**			**PCA**		**Classif.**			
	3	5	10	p = .05	p = .01	None	MW	Cohen's d	None	Yes	No	Log. Reg.	SVC	LDA	btree
S1	73.5	80.7	**81.5**	**81.5**	64.4	68.7	65.8	**81.5**	64.4	**81.5**	64.4	**81.5**	62.5	62.4	68.7
S2	**76.6**	71.3	65.2	71.0	**76.6**	71.3	-	71.0	**76.6**	**76.6**	70.2	66.2	**76.6**	61.3	70.2
S3	**79.0**	**79.0**	**79.0**	**79.0**	73.2	66.1	58.4	**79.0**	73.2	**79.0**	73.2	**79.0**	76.2	64.9	68.6
SP	**75.3**	70.8	74.0	**75.3**	73.4	74.0	-	**75.3**	70.8	**75.3**	69.2	69.2	**75.3**	65.2	65.2
SS	**66.2**	**66.2**	**66.2**	64.2	**66.2**	66.0	-	64.9	**66.2**	66.0	**66.2**	64.2	59.1	**66.2**	66.0

Table 1. Maximum UAR (%, external loop of nested CV) depending on the different elements of the classification pipeline

Ref.	Avg.	Std.	Avg. + Std.	Agg.
S1	**81.5**	63.2	62.4	68.7
S2	**76.6**	70.2	64.3	71.3
S3	**79.0**	60.0	63.7	73.2
SP	71.4	**75.3**	74.0	70.8
SS	61.8	**66.2**	65.7	66.0

Table 2. Max. UAR (%, external loop of nested CV) depending on features fusion

for syndromes, it seems that the relevant information is their variation across sessions (std). For all tasks, concatenating features over the five naps gives worse performances: we assume that this leads to excessively large feature vectors, and prevents the classifiers from generalizing – even with feature selection steps.

(ii) Decorrelation. All the systems reaching maximum performance required the use of the decorrelation block. The decorrelation threshold depends on the considered problems. While the threshold of 0.05 is best for S1 and S3, the threshold of 0.01 allows the systems classifying S2, SP, and SS to reach their maximum performances.

(iii) PCA. Regarding PCA, all the systems achieving high performances (S1, S2, S3 and SP) used PCA to reach their maximum performances. Only the system classifying SS did not use it, resulting into low performances.

(iv) Feature selection. The need to select features according to their sensitivity to sleepiness is uneven: while for S2 and SS there seems to be no need for additional feature filtering, the systems identifying S1, S3 and SP achieve better performance by selecting only features with a Cohen's d above a threshold between the *affected* and *non-affected* distributions. In all cases, filtering descriptors with a Mann-Whitney test leads to poorer performances: this statistical test is very sensitive to differences in distributions, whereas Cohen's d measures effect sizes and is less sensitive.

(v) Classification. Finally, for the classifier, the symptoms are optimally classified either with a logistic regression or with a SVC. For syndromes, the best classification performances for SP are obtained with a SVC, while for SS they are obtained with an LDA.

Estimation of syndromes from symptoms. In this section, we propose a second method for estimating syndromes from voice, by first estimating the symptoms composing the syndromes, and then merging the probabilities to estimate the syndromes. By selecting the systems achieving the best performances on S1, S2 and S3, we merge (applying an AND operator) their predictions.

This approach leads to an UAR of 70.0% for SP detection, which is lower than the performance obtained by estimating the syndrome directly from the voice (75.3%) the accumulation of errors made on each of the symptoms decreased the estimation performance of the associated syndrome. On the contrary, for the detection of SS, this approach leads to an UAR of 76.2%, which is 10% more than the direct approach. This may be due to the strong imbalance between classes in the SS detection problem (14/106 positives) while the symptoms are well classified and more balanced.

5. CONCLUSION AND PERSPECTIVES

In conclusion, we have designed a classification pipeline of three symptoms and two syndromes related to sleepiness. The results brought by this pipeline have shown the relevance of decorrelation, which not only improves the classification performance but also ensures that the selected descriptors are specific to the measured dimension with respect to the included confounding factors. Finally, we have designed classification systems estimating symptoms but also syndromes related to sleepiness, studying two different methods to estimate the latter.

Our work did not allow to settle on the best approach between direct estimation (which works better for SP) and estimation via symptoms (which works better for SS) for estimating syndromes in terms of performance. However, in terms of clinical practice, symptom-based estimation enables clinicians to better target the elements generating syndromes, and thus to customize the therapeutic formulation.

In a next step, we will study more precisely the contribution of features in classification systems – which is made possible by the fact that our pipeline is completely explicable by design. In particular, this will allow us to confirm or refute the link that seems to emerge, on the one hand, between the patients' feeling of sleepiness and acoustic descriptors; and on the other hand, between physiological sleepiness and reading quality markers (reading errors and pauses) [31].

6. REFERENCES

[1] Terry B. Young, "Epidemiology of daytime sleepiness: definitions, symptomatology, and prevalence," *The Journal of Clinical Psychiatry*, vol. 65 Suppl 16, pp. 12–16, 2004.

[2] Alexander J. Scott, et al., "Improving sleep quality leads to better mental health: A meta-analysis of randomised controlled trials," *Sleep Medicine Reviews*, vol. 60, pp. 101556, 2021.

[3] Björn Schuller, et al., "The INTERSPEECH 2011 Speaker State Challenge," in *Interspeech 2011*, 2011, pp. 3201–3204.

[4] Björn Schuller, et al., "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity," in *Interspeech 2019*, 2019.

[5] Dong-Yan Huang, et al., "Feature Normalization and Selection for Robust Speaker State Recognition," in *IEEE - International Conference on Speech Database and Assessments*, 2011.

[6] Vincent P. Martin, et al., "Sleepiness detection on read speech using simple features," in *10th Conference on Speech Technology and Human-Computer Dialogue*, 2019.

[7] Gábor Gosztolya, "Using Fisher Vector and Bag-of-Audio-Words Representations to Identify Styrian Dialects, Sleepiness, Baby & Orca Sounds," in *Interspeech 2019*, 2019, pp. 2413–2417.

[8] José Vicente Egas-López, Róbert Busa-Fekete, and Gábor Gosztolya, "On the Use of Ensemble X-Vector Embeddings for Improved Sleepiness Detection," in *Speech and Computer*, S. R. Mahadeva Prasanna, et al., Eds., 2022, Lecture Notes in Computer Science, pp. 178–187.

[9] Edward L. Campbell, et al., "Automatic detection of short-term sleepiness state. Sequence-to-Sequence modelling with global attention mechanism.," in *Workshop on Speech, Music and Mind*, 2022.

[10] Bang Tran, et al., "Speech Tasks Relevant to Sleepiness Determined With Deep Transfer Learning," in *ICASSP 2022*, 2022, pp. 6937–6941, issnlol: 2379-190X.

[11] Vincent P. Martin, et al., "Sleepiness in adults: An umbrella review of a complex construct," *Sleep Medicine Reviews*, vol. 67, pp. 101718, 2023.

[12] Christophe Gauld, et al., "A systematic analysis of ICSD-3 diagnostic criteria and proposal for further structured iteration," *Sleep Medicine Reviews*, vol. 58, pp. 101439, 2021.

[13] Geoffrey R Norman, "The epistemology of clinical reasoning: perspectives from philosophy, psychology, and neuroscience," *Academic Medicine*, vol. 75, no. 10, pp. S127–S133, 2000, publisherlol: LWW.

[14] Vincent P. Martin, et al., "Automatic Speech Recognition systems errors for objective sleepiness detection through voice," in *Interspeech 2021*, 2021, pp. 2476–2480.

[15] Vincent P. Martin, et al., "Automatic Speech Recognition system errors for accident-prone sleepiness detection through voice," in *EUSIPCO 2021*, 2021, pp. 541–545.

[16] Vincent P. Martin, et al., "Does sleepiness influence reading pauses in hypersomniac patients?," in *Speech Prosody 2022*, 2022, pp. 62–66.

[17] Vincent P. Martin, et al., "The Objective and Subjective Sleepiness Voice Corpora," in *languagelol Resources and Evaluation Conference 2020*, 2020, p. 6525-6533.

[18] Vincent P. Martin, et al., "How to Design a Relevant Corpus for Sleepiness Detection Through Voice?," *Frontiers in Digital Health*, vol. 3, pp. 686068, 2021.

[19] Donna Arand, et al., "The Clinical Use of the MSLT and MWT," *Sleep*, vol. 28, no. 1, pp. 123–144, 2005.

[20] Torbjorn Åkerstedt and Mats Gillberg, "Subjective and objective sleepiness in the active individual.," *Int J Neurosci*, vol. 52, pp. 29–37, 1990.

[21] Murray W. Johns, "A New Method for Measuring Daytime Sleepiness: The Epworth Sleepiness Scale," *Sleep*, vol. 14, no. 6, pp. 540–545, 1991.

[22] Sandeep Reddy, "Explainability and artificial intelligence in medicine," *The Lancet Digital Health*, vol. 4, no. 4, pp. e214–e215, 2022, publisherlol: Elsevier.

[23] Cynthia Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

[24] Pengcheng Guo, et al., "Recent Developments on Espnet Toolkit Boosted By Conformer," in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 5874–5878.

[25] Sylvain Galliano, Guillaume Gravier, and Laura Chaubard, "The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts," in *Interspeech 2009*, 2009, pp. 2583–2586.

[26] Vincent P. Martin, et al., "Est-il possible d'annoter la naturalité des pauses lors de la lecture d'un texte à haute voix ?," in *Journées d'Étude de la Parole 2022*, 2022.

[27] A. S. Zigmond and R. P. Snaith, "The hospital anxiety and depression scale," *Acta Psychiatrica Scandinavica*, vol. 67, no. 6, pp. 361–370, 1983.

[28] Anton Batliner, Simone Hantke, and Bjoern W. Schuller, "Ethics and Good Practice in Computational Paralinguistics," *IEEE Transactions on Affective Computing*, pp. 1–1, 2020.

[29] Russell A. Poldrack, Grace Huckins, and Gael Varoquaux, "Establishment of Best Practices for Evidence for Prediction: A Review," *JAMA psychiatry*, vol. 77, no. 5, pp. 534–540, 2020.

[30] Vincent P. Martin, et al., "How Does Comparison With Artificial Intelligence Shed Light on the Way Clinicians Reason? A Cross-Talk Perspective," *Frontiers in Psychiatry*, vol. 13, 2022.

[31] Vincent P. Martin, et al., "Physiological vs. Subjective sleepiness: what can human hearing estimate better?," in *ICPhS 2023*, 2023.