Highlights

Automatic detection of sleepiness-related symptoms and syndromes using voice and speech biomarkers

Vincent P. Martin, Jean-Luc Rouas, Pierre Philip

- Voice allows for the estimation of symptoms and syndromes related to sleepiness.
- Subjective sleepiness interferes with the acoustic quality of voice.
- Physiological sleepiness interferes preferentially with reading quality.

Automatic detection of sleepiness-related symptoms and syndromes using voice and speech biomarkers

Vincent P. Martin[a,b,c,*], Jean-Luc Rouas[b], Pierre Philip[a,d]

[a]*Deep Digital Phenotyping Research Unit, Department of Precision Health, Luxembourg Institute of Health, Strassen, L-1445, Luxembourg*
[b]*Univ. Bordeaux, LaBRI, CNRS UMR 5800, Bordeaux INP, Talence, F-33405, France*
[c]*Univ. Bordeaux, CNRS, SANPSY, UMR 6033, Bordeaux, F-33000, France*
[d]*Service Universitaire de Médecine du Sommeil, CHU de Bordeaux, Bordeaux, F-33000, France*

Abstract

Sleepiness is a major public and personal health issue. Measuring sleepiness in patients' everyday living conditions would represent a significant advancement in managing them. Voice is a signal linked to numerous health dimensions, including sleepiness. However, previous research has focused on short-term subjective sleepiness, using corpora with questionable medical validity and without measuring the specificity of identified voice biomarkers. In this article, we estimate different symptoms and long-term sleepiness-related syndromes in hypersomnolent patients. To achieve this, we have developed machine learning models that identify biomarkers that are sensitive and specific to sleepiness, reaching classification performances (Unweighted Average Recall) above 75%. Importantly, we only used statistical functions (decorrelation, Principal Component Analysis, Sensitivity test, linear classifier) so that this model remains simple and explainable to collaborating clinicians. We then leverage this explainability to identify specific vocal and speech manifestations for each type of sleepiness. By combining objective measures and the analysis of vocal characteristics, our approach provides a comprehensive understanding of long-term sleepiness and enhances patient care and management. This research holds great potential for advancing the field of digital health and contributing to improved well-being for individuals affected by sleepiness-related conditions.

Keywords:
voice biomarkers, excessive sleepiness, machine learning, automatic classification, digital health

1. Introduction

1.1. Context

Excessive sleepiness is a subject of growing scientific, social, and political attention due to its potential danger to individuals and public safety, particularly concerning road ac-

*Corresponding author
URL: vincentpmartin@protonmail.com (Vincent P. Martin)

cidents [1]. Excessive sleepiness also has consequences on individuals' health, since it is associated with an increased risk of disability and mortality related to various diseases, including sleep, metabolic, cardiovascular, neurological, and mental disorders [2, 3]. Furthermore, excessive sleepiness is often accompanied by adverse economic and social burdens, such as decreased work productivity and quality of life [4, 5, 6]. Depending on the definition of *excessive sleepiness*, its prevalence varies from 2.5% to more than 40% in the general population [7, 8, 9, 10].

1.2. The need for objective and ecological monitoring tools

While sleepiness is a natural psychophysiological state that regulates sleep and wake cycles over a 24-hour period [11, 12], excessive sleepiness arises when sleepiness occurs at inappropriate times or with increased frequency and can interfere with daily functioning [7]. When individuals report sleep disturbances, they are typically referred to a sleep specialist clinician. However, the data collected by clinicians during clinical interviews is often biased due to their own reasoning schemes [13], as well as patient biases [14] (e.g. recall biases). Therefore, clinicians require objective tools for measuring sleepiness to quantify it independently of these biases. Such a measuring tool already exists in the literature, known as the Multiple Sleep Latency Test (MSLT [15]), which involves measuring the latency of sleep onset in patients who are equipped with EEG electrodes and placed in a bed for five sessions throughout the day, with the instruction to allow themselves to fall asleep (see Part 2.1 for more details).

Yet, this procedure is costly (up to 2000€) and requires patients to be hospitalized during the test day, and the night before and after the test. This hospitalization has several other major drawbacks. Firstly, patients have to travel to the sleep center, which may not be close to their place of residence. Secondly, this hospitalization requires dedicated hospital logistics, limiting the number of tests that can be conducted. Considering the high prevalence of excessive sleepiness, there is a need for new tools that allow for the objective measurement of sleepiness on a larger scale. Furthermore, parameters measured in a single test conducted in a laboratory may differ from the patients' usual sleep behavior in their everyday living conditions, both due to environmental differences (such as noise [16]) and intrinsic variations in sleep characteristics over time [17]. Therefore, there is also a need for tools to measure sleepiness that can be implemented for regular measurement in ecological conditions, i.e., in the patients' everyday living conditions.

Given the potential for bias and the high prevalence of excessive sleepiness, detecting and monitoring sleepiness in the general population using ecologically valid methods is required to bring sleep medicine into the realm of digital medicine. A promising tool to do so is voice and speech analysis.

1.3. Voice and speech biomarkers: state of the art

Sleepiness estimation using voice and speech received less attention from the community than emblematic tasks such as depression or Parkinson's disease estimation. Nevertheless,

instantaneous sleepiness detection in speech has been the focus of two international challenges, proposed in parallel to the 2011 and 2019 Interspeech conferences[1] (respectively noted IS2011 and IS2019).

1.3.1. IS2011 challenge

During the 2011 challenge, competitors had the goal of classifying sleepiness from the Sleep Language Corpus (SLC) [18]. This corpus, extensively described in one of our previous article [19], contains the recordings of 99 healthy speakers, recorded on 9,089 samples (21 h 16 min 48 s) doing diverse speech tasks such as reading at loud a tale, sustaining vowels and reading at loud car-driver interactions. The sleepiness level is annotated using three Karolinska Sleepiness Scales (KSS) [20], which is a one-item self-questionnaire ranging from "1 - Extremely alert" to "9 - Very sleepy, great efforts to keep alert, fighting with sleepiness". One of them is filled out by the subjects themselves while the two others are completed by trained assistants [21]. Based on this annotation, competitors had to detect an average KSS score greater than 7.5, considered a 'sleepy' state [22]. This dichotomization of the data results in an unbalanced binary classification problem (34.5% of the subjects were sleepy). The metric was therefore chosen adequately to be sensitive to class imbalance: it is the Unweighted Average Recall (UAR), varying between 0 and 100%. A careful review of the six systems submitted by the competitors has been published in 2013 by Schuller et al. [23]. The winner of the challenge [24] used ASIMPLS – Asymmetric Simple Partial Least Squares [25] – and reached an UAR of 71.7%, which remains the state of the art on the SLC. It is worth noting, however, that more recent efforts have achieved a UAR of 76.4% on the SLC read-aloud subcorpus alone [26].

1.3.2. IS2019 challenge

Eight years later, "The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity" was held during the Interspeech 2019 conference [27]. For this challenge, a new corpus was introduced to the community, the SLEEP corpus (extensively described in [19]). This corpus contains the recordings of 885 healthy subjects, recorded on 10,892 samples (11h 41 min 17 s). The recording tasks are unknown, and the data have been annotated with the rounded value of the same annotation as performed on the SLC (average of three KSS). The large size of this corpus made it possible to introduce a new automatic learning task through the estimation of the severity of the speakers' sleepiness (i.e. the KSS score), whose performance is measured by the Pearson correlation between the predictions and the ground truth values. On the six competitors, three used classical machine learning models [28, 29, 30] while three introduced deep-learning-based techniques to this task [31, 32, 33]. However, the best approach to this day is based on extracting Fischer vectors from the spectrum and estimating sleepiness using a Support Vector Regressor (SVR), leading to a Pearson correlation coefficient of $r = .387$ [30]. This score has never been outreached since then, even by systems using the

[1] The Interspeech conference is the flagship conference of the International Speech Communication Association (ISCA).

latest deep learning techniques. Indeed, Fritsch et al. reached $r = .325$ using end-to-end convolutional neural networks [34]; Amiriparian et al. obtained $r = .367$ using the fusion of attention and non-attention autoencoders [35]; Egas-Lopez et al. reached $r = .365$ by using x-vectors and SVR [36, 37]. The most recent approach, based on a Sequence-to-Sequence model with global attention mechanism, achieved $r = .383$ in [38].

1.3.3. Voiceome dataset

During the international conference ICASSP 2022, a new large-scale dataset was introduced: the Voiceome dataset, containing the recordings of 1828 subjects (186.2 h) from the general population. These recordings were made by the subjects using their smartphone, in ecological conditions, on 11 different tasks comprising spontaneous speech, reading out loud, sustained phonation, and diadochokinetic tasks. They have been annotated by the subject themselves using the Standford Sleepiness Scale (SSS) [39], a one-item scale ranging from '1 - Feeling active, vital, alert, or wide awake' to '7 - No longer struggling to sleep, sleep onset early, having dream-like thoughts'. The proposed system, based on masking and separated training, reached an accuracy of 81.13% [40].

1.3.4. Multiple Sleep Latency Test corpus

In parallel of these works on instantaneous subjective sleepiness, we designed and collected a corpus – the Multiple Sleep Latency Test corpus [41, 19] – at the Sleep Clinic of the Bordeaux University Hospital, intending to estimate the objective long-term sleepiness of hypersomniac patients. We previously used a combination of acoustic and Automatic Speech Recognition system errors to detect with good classification performances objective long-term sleepiness (73.2% of UAR in [42]) and subjective long-term sleepiness (74.2% of UAR in [43]).

1.4. Limits

The limitations of the previously published work are twofold. First, neither of the two scales used to annotate the data in the SLC, the SLEEP corpus and the Voiceome dataset (KSS or SSS) is used in clinical practices in sleep medicine [44, 45]. While the average of the three KSS used for the SLC and the SLEEP corpus does not exist at all in the sleep medicine literature, the SSS suffers from the bidimensionality of the construct it measures [46]. This hinders the use and endorsement of these tools by clinicians, who discern numerous psychophysiological constructs around sleepiness [47]. Moreover, recent perceptual experiments discussed the validity of the SLEEP corpus for sleepiness detection using voice recordings [48, 49].

Moreover, "the clinician attempting to make a diagnosis is dealing almost exclusively at the syndrome level" [50], a *syndrome* being defined as "a cluster of signs and symptoms" [50] 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 usually investigate several symptoms and syndromes related to sleepiness. While some of our previous works have initiated a multidimensional study of long-term sleepiness in sleep clinic patients, our objective is to go further and propose the classification of both symptoms and syndromes related to excessive sleepiness.

1.5. Objectives

In this study, we aim to bring the machine learning problem formulation of sleepiness detection through voice closer to clinical reasoning reality. In this way, we propose to investigate the classification of three long-term sleepiness-related symptoms and two corresponding syndromes, that are defined as different combinations of symptoms. Contrary to previous works focusing on the estimation of short-term sleepiness, we focus in this article only on long-term sleepiness. We consider two methods to estimate syndromes: either directly from voice biomarkers, using the same classification model as symptoms, but with the database labeled adequately; or combining the estimation of symptoms defining each syndrome to estimate syndromes directly from voice recordings.

In order to achieve this goal, we extend our previous work on the Multiple Sleep Latency Test corpus [19, 41] on objective [42] and subjective [43] measures of excessive sleepiness. In addition to acoustic features and the errors made by Automatic Speech Recognition systems, we aim at measuring the contribution of reading pauses (number, duration but also location in the read text [51]) for the detection of excessive sleepiness using speech. Moreover, we aim to design *biomarkers* of these symptoms and syndromes, thus features that are both *sensible* and *specific* to sleepiness. Finally, since we work in close collaboration with sleep clinicians, we require our features and model to be explainable by design [52] so that clinicians can understand and be confident about the proposed system [53].

2. Method

2.1. Multiple Sleep Latency Corpus – MSLTc

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 [41, 19]. They undertake a Multiple Sleep Latency Test (MSLT), consisting of asking them to take 20-minute 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 [15].

Before each nap, 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 (KSS) [20]. During their stay at the hospital, the patients are asked to fill out numerous sleep-related and health questionnaires [41], including the Epworth Sleepiness Scale [34, ESS], a questionnaire measuring the subjective perception of sleep propensity during the two previous weeks.

2.2. Symptom and syndrome definitions

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 three symptoms and two syndromes, representing five different definitions of sleepiness. They are represented in the Figure 1.

The considered symptoms are the following:

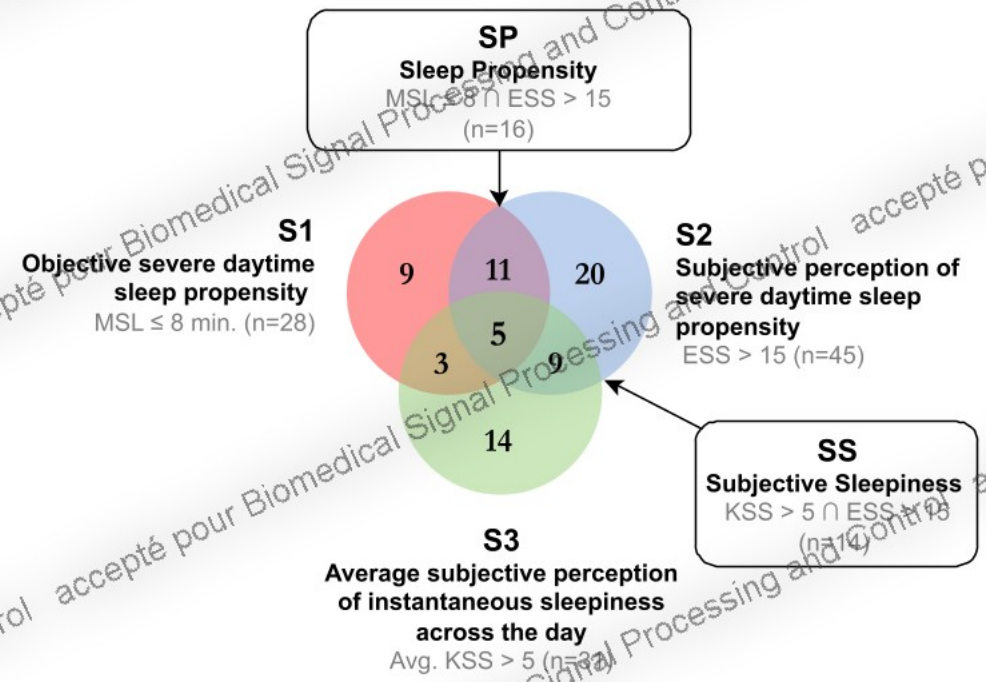

Figure 1: Venn diagram of the symptomatic profiles of the patients in the MSLTc

(S1) *Objective severe daytime sleep propensity*, measured by an MSL $\leq$ 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, 42.4%).

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

From all these symptomatic profiles, two syndromes are of particular interest for sleep clinicians [47]:

(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 $\leq$ 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 ∩ S3). In this binary classification task, 14/106 patients (13.2%) are affected by the SS syndrome.

2.3. Voice and speech features

All the features that we extracted from the voice recordings and used as input to our classification model are listed in Table 1.

Family	Type	Metric	Abbreviation
Acoustic	F0	mean	F0mean
		variance	F0var
		slope	F0slope
	Energy	mean	NRJmean
		variance	NRJvar
		slope	NRJslope
	Formant {1,2,3,4}	Frequency	Formant F{1,2,3,4}
		Bandwidth	Formant bw{1,2,3,4}
		Amplitude	A{1,2,3}
	Harmonics {1,2,4}	Amplitude	H{1,2,4}
	H2 - H1, H4 - H2		H1H2, H2H4
	A{1,2,4} - H1		A{1,2,4}H1
	Harmonic-to-noise ratio	[0-500Hz]	HNR1
		[0-1500Hz]	HNR2
		[0-2500 Hz]	HNR3
		[0-3500Hz]	HNR4
	Cepstral Peak Prominence		CPP
ASR errors	Insertion	Number	WIns
		Ratio	%WIns
	Deletion	Number	WDel
		Ratio	%WDel
	Substitution	Number	WSub
		Ratio	%WDel
	Correct	Number	WCorrect
		Ratio	%WCorrect
Reading pauses	{Correct, Incorrect, Total}	Number of pauses	Nb {+, -, Tot}
		Nb/N_Tot	Ratio_{+,-}
		Total time	T_{+,-,Tot.}
		Naturalness score	S_{+, -, Tot}
		Time-weighted naturalness score	WS_{+, -, Tot}

Table 1: Features automatically extracted from audio recordings

Constraints. As this work is carried out in close collaboration with sleep clinicians, we set the interpretability of our results as a major criterion of our work [53]. We therefore decided to put this factor as a priority in the design of our descriptors and our classification model [52]. Therefore, we used descriptors that are understandable by non-specialists in signal processing and designed our machine learning classification model with functions to which they are already familiar (PCA, logistic regression, ...). Three families of descriptors are extracted from the audio recordings.

Acoustic parameters. First, we extracted features related to the quality of the voice ($n = 30$) [26, 55] that we validated with clinicians. After having automatically identified the voiced parts in recordings, we estimated the fundamental frequency, energy, Harmonic-to-Noise Ratio, ... using the `Snack` toolbox [56].

Automatic Speech Recognition errors ($n = 112$). The second set of features used in this study is intended to complement the previous information on acoustic voice quality with metrics related to reading quality, reflecting the interference of sleepiness with the neurolinguistic processes involved in reading aloud. In a preliminary study [57], we have shown the efficiency of hand-labeled reading mistakes for the detection of objective daytime sleep propensity ($MSL \leq 8$, Symptom n°1). However, this manual annotation is very costly both in terms of time and expertise and is not compatible with the goal of a fully automated model. Therefore, we sought to automate the extraction of these reading errors by studying the errors made by automatic speech recognition systems. We consider in this study four descriptors (Insertions, Substitutions, Deletions, and the number of correctly identified tokens Correct) made by a conformer-based end-to-end ASR system, implemented in `Espnet` [58] and trained on ESTER [59]. The errors are computed on words, and we consider both the number of errors and their ratio over the total number of identified words. As our ASR system is an end2end system using Conformer blocks, this sequence-to-sequence design does not allow to keep the time alignements between frames and output symbols.

These features have previously been proven useful in the detection of Symptom n°1 ($MSL \leq 8$ min) [42] and Symptom n°2 ($ESS > 15$) [43]. They are therefore promising for the detection of the syndromes targeted in this article.

Pauses duration and location ($n = 14$). To complement the estimate of reading ability given by reading errors, we studied reading pauses: their number, their duration but also their location in the text, using a combination of ASR system and a Voice Activity Detector. Then, thanks to annotations made by speech therapists, we were able to estimate whether the reader stops at "natural" locations (e.g. at the end of a sentence) or "unnatural" ones (e.g. in the middle of a sentence) [60]. Using this information, we computed features derived from the number of pauses, their duration, naturalness scores, and duration-weighted naturalness scores. These features have been previously linked to both objective sleepiness (Symptom n°1) and subjective sleepiness (Symptom n°2 and Symptom n°3). We redirect the reader to a previous publication for more details [51].

2.4. Classification model

Our classification model is represented in Figure 2. In the same manner as the features, it has been designed to be understood and has been validated by clinicians. We trained and evaluated five models, each corresponding to the classification of a symptom or a syndrome related to sleepiness.

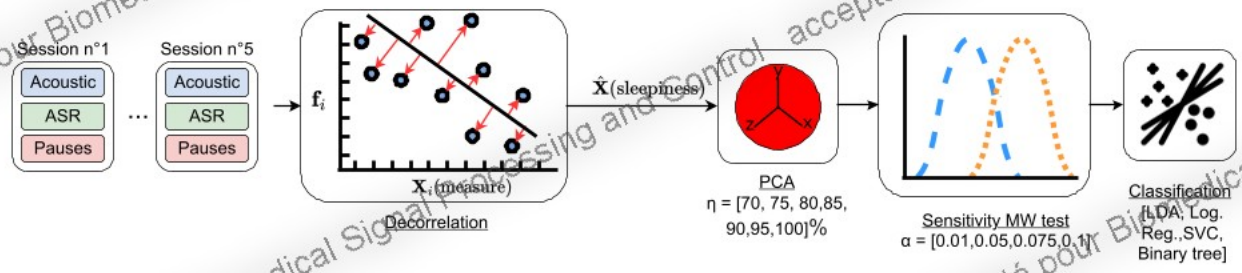

Figure 2: Classification model and hyperparameters.
ASR: Automatic Speech Recognition errors, f_i: confounding factors, MW: Mann-Whitney, PCA: Principal Component Analysis, LDA: Linear Discriminant Analysis, Log. Reg.: Logistic Regression, SVC: Support Vector Classifier.

Features averaging and aggregation. We aim to detect long-term sleepiness-related symptoms and syndromes from the five recordings collected during a MSLT. To do so, we tested four feature fusion strategies: averaging them across the naps, keeping the average and the standard deviation, keeping only the standard deviation, and aggregating all features. Moreover, to limit the features to the essential minimum, we tested all different combinations of feature categories (Acoustic, ASR errors, and Pauses). Features are then fed into the following pipeline.

(i) Decorrelation (specificity). First, a decorrelation algorithm is applied to the features. 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 [61]), and educational level (as measured by the highest level obtained after the French Certificate of Education).

To decorrelate each feature $\mathbf{X}_i$(measured), we assume that it can be written as the combination of a contribution from sleepiness $\hat{\mathbf{X}}_i$(sleepiness) and a contribution from the confounding factors correlating with $\mathbf{X}_i$(measured), denoted by $\hat{\mathbf{X}}_i(\mathbf{f}_i)$:

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

The decorrelation process is thus the following:

1. For each measured feature $\mathbf{X}_i$(measured), we identified the confounding factors correlating with it (Pearson's test p-value $p < \beta \in \{.05, .01\}$), denoted as $\mathbf{f}_i$.

2. We estimate the contribution of the confounding factors $\mathbf{f}_i$ to $\mathbf{X}_i$(measured), noted by $\hat{\mathbf{X}}_i(\mathbf{f}_i)$, by a multivariate linear regression, i.e. $\hat{\mathbf{X}}_i(\mathbf{f}_i) = \sum_{f_i} \alpha_i f_i$
3. We finally estimate $\hat{\mathbf{X}}_i$(sleepiness), the contribution of sleepiness to $\mathbf{X}_i$(measured), by:

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

The features still correlating Pearson's test p-value ($p < \beta$) with at least one confounding factor after this procedure are excluded from the selected features since they are not specific to sleepiness regarding the measured confounding factors. An example of such a decorrelation process is given in Annex Appendix B.

(ii) PCA. Next, to reduce the dimensionality, to orthogonalize and to maximize the variance in each dimension of the selected features, we perform a Principal Component Analysis (PCA), keeping $\eta\% \in \{75, 80, 85, 90, 95\}$ of the original variance.

(iii) Sensitivity to sleepiness. Then, we filter out PCA dimensions that are not sensitive enough for sleepiness, in order to select only *biomarkers* of sleepiness. To do so, we select dimensions for which either a Mann-Whitney test ran on the *affected* vs. *non-affected* classes gives a p-value lower than $\alpha \in \{.01, .05, .075, .1\}$ (using the `SelectKBest` function of the `scikit-learn` Python package [62]), or those having the $k\% \in \{5, 10, 25, 50, 75, 90\}$ best Cohen's d (using `SelectPercentile`).

(iv) Classification. Finally, the classification is performed by a Support Vector Classifier (SVC) with a linear kernel, a logistic regression, a Linear Discrimant Analysis (LDA) or a binary tree, all known by clinicians and allowing the explainability of the selected descriptors. Since the classification problems are unbalanced, weights are biased by the number of positive samples in each class during the training.

Nested cross-validation and performance metric. In line with community guidelines [63], performances are computed using the Unweighted Average Recall (UAR), a relevant metric for unbalanced binary classification problems, defined as the average of the recall on the positive class and recall on the negative class.

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 UAR on all the predictions vs. ground truth, to limit approximations due to averaging metrics computed on few samples.

While LOSO has been described as inflating performances on regression tasks [64], it is used in the external loop to have the same training and test bases for each of the symptoms and syndromes, in order to provide valid comparison. Doing so, each testing sample corresponds to the aggregated features of one patient, mimicking the clinical reasoning, i.e. evaluating one patient based on the knowledge accumulated on the previous ones [13].

2.5. Other classification tasks

Estimation of syndromes from symptoms. In addition to the detection of three sleepiness-related symptoms and two syndromes from speech recordings, we also evaluated a second method for estimating syndromes, 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 merged (applying an AND operator) their predictions: $y^{SP}_{pred} = y^{S1}_{pred} \cap y^{S2}_{pred}$ and $y^{SS}_{pred} = y^{S2}_{pred} \cap y^{S3}_{pred}$.

Differential diagnosis. To ensure that the classifiers have learned the desired task, we also evaluated the performances of each classifier to detect the other symptoms and syndromes that it has not been trained to identify, by comparing the predictions made by each classifier with the ground truths of other symptoms or syndromes. These performances give a measure of the specificity of the prediction obtained by a classifier to the symptom or syndrome it has been trained on: the more specific a classifier is for one symptom or syndrome, the lower its prediction score should be for another symptom, and vice versa.

2.6. Contribution of the vocal features implied in each classifier's decision

In order to identify the voice parameters that have the strongest weights in the final classification, we computed, for each LOSO external loop iteration and each descriptor, the product of the weight of this descriptor in the PCA and the weight of each PCA component in the classification. Then, we computed the absolute average contribution of each descriptor across the LOSO.

3. Results

The classification results of symptoms and syndromes based on recordings are presented in Table 2, while the results of the differential diagnosis are presented in Table 3. Finally, the analysis of audio descriptors contributing to the classification is provided in Section 3.3.

3.1. Classification model performances

Task	UAR	k_{inner}	Fusion	Combination	Decorrelation	Sen.	PCA	Classif.
S1	81.5%	10	Avg.	All	$p = .05$	Cohen	Yes	Log. Reg.
S2	76.6%	3	Avg.	Ac.	$p = .01$	None	Yes	SVC
S3	79.0%	3, 5, 10	Avg.	Ac.	$p = .05$	Cohen	Yes	Log. Reg.
SP	75.3%	3	Std.	Ac.	$p = .05$	Cohen	Yes	SVC
SS	66.2%	3, 5, 10	Std.	Ac.	$p = 0.01$	None	No	LDA

Table 2: Best hyperparameter combination for each classifier and the corresponding UAR.
k_{inner}: number of folds in the inner cross-validation loop, *Fusion*: function for the aggregation of the 5 sets of features/patient, *Combination*: features combination, *Ac.*: Acoustic features only, *Decorrelation*: p-value threshold for the decorrelation process (see Appendix B), *Sen.*: Sensibility function

Model performances. Our automatic classification pipeline achieve Unweighted Average Recalls(UAR) of respectively 81.5%, 76.6%, and 79.0% for the detection of S1, S2, and S3. However, the classification performance is less consistent for syndromes: while Sleep Propensity (SP) is detected with a UAR of 75.3%, the classification of Subjective Sleepiness (SS) achieves only an UAR of 66.2%. Performances of each pipeline depending on its hyperparameters are reported in Appendix C.

Estimation of syndromes from symptoms. Combining the predictions of the two symptoms composing each syndrome a posteriori results in an UAR of 70.0% for estimating SP, which is slightly lower than the direct estimation (75.3%): the accumulation of errors made on the estimation of each symptom decreases the performance of the associated syndrome. On the other hand, for SS, combining the estimations of the two symptoms S2 and S3 yields an UAR of 76.2%, which is 10% higher than the pipeline estimating this syndrome directly from voice. One possible cause may be the highly imbalanced class distribution for the subjective sleepiness detection problem (14/106 positives), while the symptoms are slightly more balanced (45/106 positives for S2, 31/106 for S3), making it easier to build classifiers.

3.2. Differential diagnosis

		prediction				
		S1	S2	S3	SP	SS
ground truth	S1	**81.5**	59.2	49.8	70.7	51.7
	S2	51.7	**76.6**	52.3	59.2	64.5
	S3	47.5	55.4	**79.0**	50.1	64.0
	SP	**73.1**	66.4	48.8	**75.3**	57.3
	SS	45.8	**80.4**	64.2	57.2	**66.2**

Table 3: Performance of each model on every task. For instance, the estimation of S1 using the ground truth of SP yields an UAR of 73.1%.

The performance of each model for every task is reported in Table 3. The diagonal of this table corresponds to the performances reported in Table 2. The higher the performance, the better the system has generalized the task on which it was trained.

The other performances correspond to the estimation of symptoms and syndromes other than those on which the models have been trained. Thus, the system trained on symptom S1 achieves an UAR of 70.7% for predicting syndrome SP, and the system trained on syndrome SS estimates symptom S2 with an UAR of 80.4%, outperforming the specifically trained S2 system for this task.

3.3. Voice features implied in the classification systems

The feature contribution to each symptom classification task is plotted in Figure 3, while the contribution of features for the classification of syndromes is represented in Figure 4. For each of these figures, the green color refers to pause-related features, while the orange color refers to acoustic features. The angle of each disc portion is proportional to the

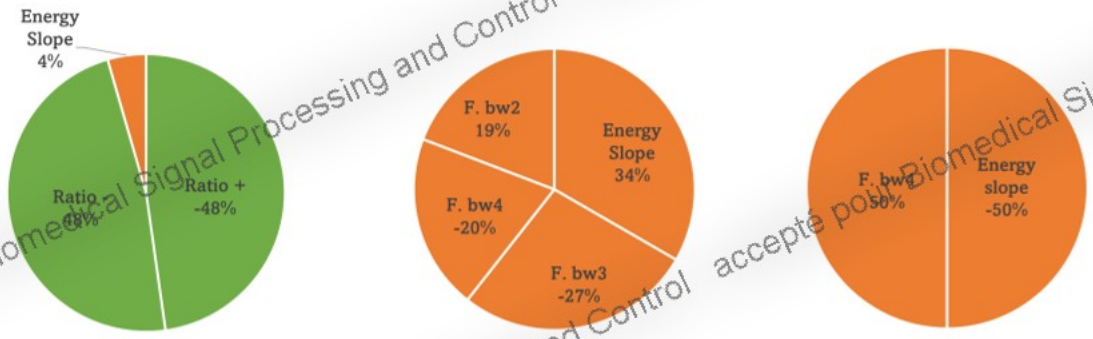

Figure 3: Voice and speech features implied in the classification of the objective severe daytime sleep propensity (S1, left); subjective perception of severe daytime sleep propensity (S2, middle); and average subjective perception of instantaneous sleepiness across the day (S3, left).

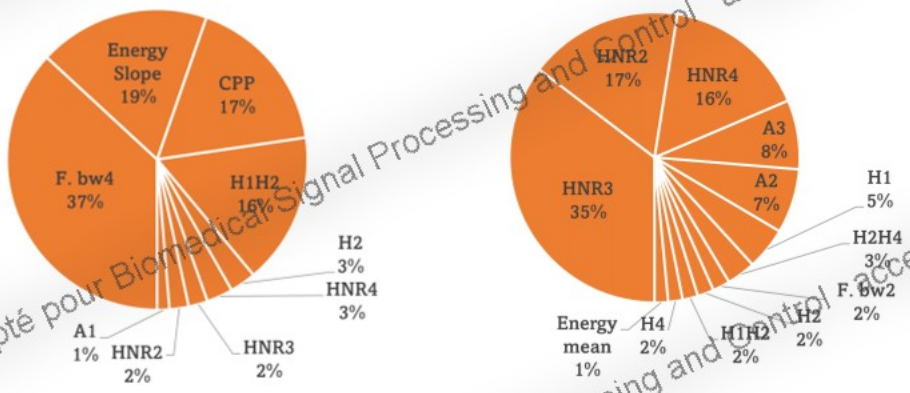

Figure 4: Voice and speech features implied in the classification of sleep propensity (SP, left) and subjective sleepiness (SS, right)

contributions of each descriptor involved in the classification. The numbers in parentheses are the relative contributions of each descriptor. An example of such feature values on two patients (one having all the three symptoms and another having none of them) has been reported in Appendix A. More detailed metrics on the contribution of each descriptor to the classification are provided in Table D.8 to D.12 in the Appendix D.

The classifier detecting objective sleep propensity (S1) is the only one relying on features related to pauses, namely the ration of correct and incorrect pauses. The third feature is related to acoustic energy. Thus, a patient with symptom S1 will be identifiable not only by a higher energy slope but above all by the less natural placement of reading pauses.

For the detection of both S2 and S3 symptoms, both the energy slope and the bandwidth characteristics of the formants are important in the classification, but with opposite effects. While patients suffering from S2 are identified by a greater energy slope and a decrease in formant 4 bandwidth, patients suffering from S3 are on the contrary recognizable by their lower energy slope and increased formant 4 bandwidth.

Concerning syndromes, their identification mobilizes many more different vocal features, reflecting the complexity of the task compared to the identification of symptoms. Thus, sleep propensity (SP) is identified by an increase in 4th formant bandwidth, Cepstral Peak Prominence, slope energy and the difference between H1 and H2; subjective sleepiness (SS) is identifiable by its alteration in HNR and 2nd and 3rd formant amplitude.

4. Discussion

The contribution of this article is twofold: on the one hand, we investigate the link between various sleepiness-related symptoms and changes in the voice and speech of hypersomnia patients; these links are discussed in Section 4.1. Secondly, we are building machine-learning models of different conceptual levels of clinical reasoning: symptoms and syndromes. The conceptual level learned by each model is discussed in Section 4.2.

4.1. Impact of physiological and subjective sleepiness on voice

In their 2009 seminal paper, Krajewski et al. [22] identified five pathways of the impact of sleepiness on speech production:

- Cognitive speech planning: reduced cognitive processing speed → slackened articulation and slowed speech ;
- Respiration: decreased of muscle tension → lower fundamental frequency, intensity, articulatory precision, and rate of articulation ;
- Phonation: decreased muscle tension → decreased resonance frequencies (formants) positions and broadened formant bandwidth.
- Articulation: decreased muscle tension → shift in the spectral energy distribution, broader formant bandwidth, increase in formant frequencies especially in lower formants.
- Radiation: decreased orofacial movement → broadened Formant 1 bandwidth, smaller Formant 1 amplitude.

However, these pathways have been proposed only for instantaneous subjective sleepiness. Our results complete this model with the impact of physiological (S1), long-term subjective (S2) and average sleepiness (S3) on speech production.

First, the expression through voice and speech of average sleepiness over the day (S3) is in agreement with the previous model: a higher level of average sleepiness is associated with an increase in the bandwidth of the 4th formant and a decrease in slope energy. These descriptors have previously been associated with phonation and articulation effects for the bandwidth of the 4th formant, and with breathing-related factors for the decrease in the energy slope. The correspondence between our results and the model of Krajewski et al. is explained by the fact that the S3 symptom is the average of several estimates of instantaneous sleepiness, reflecting the average of the phenomena across the recording sessions.

However, this has to be distinguished from the impact of subjective perception of sleep propensity (S2), which is a single subjective assessment of sleepiness over the last few weeks. Indeed, not only does this symptom affect the bandwidth of formants 2 (which increases) and 3 (which decreases), but the coefficients attributed to the slope energy and the bandwidth of formant 4 are inverted.

Finally, thanks to the study of symptom S1, we add a new dimension to the previous model by revealing the specific impact of physiological sleepiness on speech production. Indeed, whereas for healthy subjects physiological sleepiness and subjective sleepiness correlate [65], the subjects we recorded were diagnosed as hypersomniac, a disease for which perception does not necessarily correspond to physiological sleepiness [66], allowing us to study each separately. Thus, our results show that physiological sleepiness does not have a major impact on acoustic quality, as is the case for the two subjective sleepiness symptoms, but on reading quality, and more specifically on the location of reading pauses. Thus, the higher the physiological sleepiness of the patients (i.e., low sleep latency), the more they tend to distribute their reading pauses in locations considered unnatural [60]. Acoustic descriptors have only a small contribution to the classification of symptom S1, explaining our previous failures to classify physiological sleepiness based solely on acoustic descriptors [67]. Moreover, this strong difference between subjective sleepiness, which impacts acoustic voice quality, and physiological sleepiness, which impacts reading quality, is consistent with a recent perceptual study on the difference in physiological and subjective sleepiness perception [68].

4.2. What concept has been generalized by each system?

In a second step, we aim at discussing the concept that has been generalized by each system. Indeed, when training a system to reproduce the ground truth of a database, it is generally accepted that the classification system has generalized the concept contained in the database annotation [69, 70, 71, 72, 73]. We propose here to discuss the conceptual level learned by the classifiers: did they generalize the concept presented to them, or on the contrary a higher-level concept (e.g. the syndrome instead of the symptom) or a lower-level concept (e.g. a symptom instead of the syndrome)?

Based on the results of the differential diagnosis, the predictions made by the S1 classifier allow the classification of both S1 symptoms (UAR = 81.5%) and, to a lesser extent, the SP syndrome (UAR = 73.1%). Therefore, we hypothesize that the classifier trained to estimate S1 has generalized the concept of an objective measure of the propensity to fall asleep during the day, which allows the partial detection of the sleep proportion syndrome. In return, the estimates made by the classifier trained to detect the SP syndrome allow detecting the SP syndrome with an UAR of 75.3% and the S1 symptom with an UAR of 70.7%, but with lower performances concerning the S2 symptom which also composes the syndrome. This seems to indicate that the expression of sleep propensity (SP) in the voice manifests itself more closely to the objective sleep propensity (S1) than to the subjective sleep propensity (S2).

On the contrary, the classifier trained to estimate S2 allows estimating the SS syndrome with more precision than the system trained to do so: we assume this system has learned a

more general concept (i.e. subjective sleepiness) than the one it was trained on (i.e. subjective perception of severe daytime sleep propensity, S2). This is in line with the observation that the classifier trained to estimate S3 stands out of the crowd and has very specifically learned the concept of average subjective sleepiness (S3).

In conclusion, our results do not allow us to decide between a direct estimate of syndromes and a primary estimate of symptoms. On the other hand, for clinical use, symptom estimation is a much more flexible option, allowing the clinician to follow the desired symptoms, and even recompose them into syndromes.

5. Conclusion and Perspectives

This study paves the way for the integration of sleepiness measurement tools directly with patients in a medical context, using voice recordings. This disruptive approach will address the limitations of the MSLT concerning hospital noise, natural variations in EEG parameters, high cost of the test while providing crucial information to clinicians about the progression of patients' symptoms in their everyday living environment. This approach allows, for example, measuring the therapeutic efficacy of proposed treatments or early detection of relapse. To achieve this, we have made technological and scientific contributions to the problem of designing vocal biomarkers for sleepiness [74].

Firstly, we have designed a fully explainable automatic classification model that enables the detection of three symptoms and two syndromes related to long-term sleepiness in hypersomnolent patients, with performance exceeding 75% of UAR. Both this classification model and the descriptors extracted from the voice recordings were intentionally kept as simple as possible to facilitate dialogue with the collaborating clinicians. Additionally, particular attention was given to the features selection step, aiming for descriptors that are sensitive but also specific to sleepiness. Specificity was ensured through the introduction of a decorrelation step in the model involving seven confounding factors.

Secondly, the scientific contribution of our work has been to leverage the explainability of our classification process to study the link between different symptoms and syndromes related to long-term sleepiness and their manifestation in speech. This complements previous studies that have focused on short-term subjective sleepiness. Our results seem to indicate a specific connection between physiological sleepiness and decreased reading performance (misplaced pauses). Furthermore, a strong relationship was observed between the perception of sleepiness and the acoustic quality of the voice (energy, formant bandwidth). However, these preliminary results need to be confirmed by further experimental studies.

Our future work aims to expand upon these initial results obtained in highly controlled conditions (hospital environment, high-quality microphone, reading aloud tasks) to more naturalistic conditions. To achieve this, data collection is already underway, involving recordings of hypersomnolent patients using smartphones in tasks such as reading aloud, semi-spontaneous speech (e.g., describing the route to this location), and spontaneous speech relevant to physicians (e.g., "Do you feel like it takes a long time for you to get up when you wake up in the morning?"). These new tasks will require the adaptation of certain speech recording features (e.g., the placement of reading pauses) or the exploration of new

dimensions in speech production, such as the phonetic aspects that we have already begun to study [75]. Furthermore, data collection is currently underway on healthy subjects undergoing sleep deprivation. This will enable us to compare the expression of different types of sleepiness in the voices of both healthy subjects and hypersomnolent patients.

6. Acknowledgements

This study has been partially funded by the SOMVOICE project, sponsored by the Labex BRAIN (ANR-10-LABX-43). VPM has received funding from the European Union's Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No. 101106577.

References

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

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

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

[4] C. M. Barnes, N. F. Watson, Why healthy sleep is good for business, Sleep Medicine Reviews 47 (2019) 112–118. URL: https://www.sciencedirect.com/science/article/pii/S1087079219300449. doi:10.1016/j.smrv.2019.07.005.

[5] D. Léger, C. Stepnowsky, The economic and societal burden of excessive daytime sleepiness in patients with obstructive sleep apnea, Sleep Medicine Reviews 51 (2020) 101275. URL: https://www.sciencedirect.com/science/article/pii/S1087079220300186. doi:10.1016/j.smrv.2020.101275.

[6] D. Léger, V. Bayon, J. P. Laaban, P. Philip, Impact of sleep apnea on economics, Sleep Medicine Reviews 16 (2012) 455 462. URL: https://www.sciencedirect.com/science/article/pii/S1087079211000992. doi:10.1016/j.smrv.2011.10.001.

[7] M. M. Ohayon, Operational Definitions and Algorithms for Excessive Sleepiness in the General Population: Implications for DSM-5 Nosology, Archives of General Psychiatry 69 (2012) 71. URL: http://archpsyc.jamanetwork.com/article.aspx?doi=10.1001/archgenpsychiatry.2011.1240. doi:10.1001/archgenpsychiatry.2011.1240.

[8] I. Jaussent, C. M. Morin, H. Ivers, Y. Dauvilliers, Incidence, worsening and risk factors of daytime sleepiness in a population-based 5-year longitudinal study, Scientific Reports 7 (2017) 1372. URL: https://www.nature.com/articles/s41598-017-01547-0. doi:10.1038/s41598-017-01547-0.

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

[10] B. P. Kolla, J.-P. He, M. P. Mansukhani, M. A. Frye, K. Merikangas, Excessive sleepiness and associated symptoms in the U.S. adult population: prevalence, correlates, and comorbidity, Sleep Health 6 (2020) 79–87. doi:10.1016/j.sleh.2019.09.004.

[11] J. Shen, J. Barbera, C. M. Shapiro, Distinguishing sleepiness and fatigue: focus on definition and measurement, Sleep Medicine Reviews 10 (2006) 63–76. URL: https://linkinghub.elsevier.com/retrieve/pii/S1087079205000444. doi:10.1016/j.smrv.2005.05.004.

[12] M. M. Ohayon, From wakefulness to excessive sleepiness: What we know and still need to know, Sleep Medicine Reviews 12 (2008) 129 141. URL: https://www.sciencedirect.com/science/article/pii/S1087079208000075. doi:10.1016/j.smrv.2008.01.001.

[13] V. P. Martin, J.-L. Rouas, P. Philip, P. Fourneret, J.-A. Micoulaud-Franchi, C. Gauld, How Does Comparison With Artificial Intelligence Shed Light on the Way Clinicians Reason? A Cross-Talk Perspective, Frontiers in Psychiatry 13 (2022). doi:10.3389/fpsyt.2022.926286.

[14] S. Mouchabac, I. Conejero, C. Lakhlifi, I. Msellek, L. Malandain, V. Adrien, F. Ferreri, B. Millet, O. Bonnot, A. Bourla, R. Maatoug, Improving clinical decision-making in psychiatry: implementation of digital phenotyping could mitigate the influence of patient's and practitioner's individual cognitive biases, Dialogues in Clinical Neuroscience 23 (2021) 52 61. URL: https://www.tandfonline.com/doi/full/10.1080/19585969.2022.2042165. doi:10.1080/19585969.2022.2042165.

[15] D. Arand, M. Bonnet, T. Hurwitz, M. Mitler, R. Rosa, R. B. Sangal, The Clinical Use of the MSLT and MWT, Sleep 28 (2005) 123–144. doi:10.1093/sleep/28.1.123.

[16] E. De Lima Andrade, D. C. Da Cunha E Silva, E. A. De Lima, R. A. De Oliveira, P. H. T. Zannin, A. C. G. Martins, Environmental noise in hospitals: a systematic review, Environmental Science and Pollution Research 28 (2021) 19629–19642. URL: https://link.springer.com/10.1007/s11356-021-13211-2. doi:10.1007/s11356-021-13211-2.

[17] A. Chouraki, J. Tournant, P. Arnal, J.-L. Pépin, S. Bailly, Objective multi-night sleep monitoring at home: variability of sleep parameters between nights and implications for the reliability of sleep assessment in clinical trials, SLEEP 46 (2023) zsac319. URL: https://academic.oup.com/sleep/article/doi/10.1093/sleep/zsac319/6965459. doi:10.1093/sleep/zsac319.

[18] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, The INTERSPEECH 2011 Speaker State Challenge, in: Interspeech 2011, 2011, pp. 3201–3204. doi:10.1.1.364.4935.

[19] V. P. Martin, J.-L. Rouas, J.-A. Micoulaud-Franchi, P. Philip, J. Krajewski, How to Design a Relevant Corpus for Sleepiness Detection Through Voice?, Frontiers in Digital Health 3 (2021) 686068. URL: https://www.frontiersin.org/articles/10.3389/fdgth.2021.686068/full. doi:10.3389/fdgth.2021.686068.

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

[21] M. Golz, D. Sommer, M. Chen, D. Mandic, U. Trutschel, Feature Fusion for the Detection of Microsleep Events, Journal of VLSI Signal Processing 49 (2007) 329 342.

[22] J. Krajewski, A. Batliner, M. Golz, Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach, Behavior Research Methods 41 (2009) 795–804.

[23] B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, F. Weninger, F. Eyben, Medium-term speaker states-A review on intoxication, sleepiness and the first challenge, Comput. Speech Lang. 28 (2013) 346–374.

[24] D.-Y. Huang, S. S. Ge, Z. Zhang, Speaker State Classification Based on Fusion of Asymmetric SIMPLS and Support Vector Machines, in: Interspeech 2011, 2011, p. 4.

[25] S. DeJong, SIMPLS: An alternative approach to partial least squares regression, Chemometrics and Intelligent Laboratory Systems 18 (1993) 251 263.

[26] V. P. Martin, J.-L. Rouas, P. Thivel, J. Krajewski, Sleepiness detection on read speech using simple features, in: 10th Conference on Speech Technology and Human-Computer Dialogue, Timisoara, Romania, 2019. doi:10.1109/SPED.2019.8906577.

[27] B. Schuller, A. Batliner, C. Bergler, F. B. Pokorny, J. Krajewski, M. Cychocz, R. Vollman, S.-D. Roelen, S. Schnieder, E. Bergelson, A. Cristia, A. Seidl, A. Warlaumont, L. Yankowitz, E. Nöth, S. Amiriparian, S. Hantke, M. Schmitt, The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity, in: Interspeech 2019, 2019. doi:10.21437/Interspeech.2019-1122.

[28] V. Ravi, S. J. Park, A. Afshan, A. Alwan, Voice Quality and Between-Frame Entropy for Sleepiness Estimation, in: Interspeech 2019, 2019, pp. 2408 2412. URL: http://www.isca-speech.org/archive/Interspeech_2019/abstracts/2988.html. doi:10.21437/Interspeech.2019-2988.

[29] H. Wu, W. Wang, M. Li, The DKU-LENOVO Systems for the INTERSPEECH 2019 Computational Paralinguistic Challenge, in: Interspeech 2019, 2019, pp. 2433–2437. URL: http://www.isca-speech.org/archive/Interspeech_2019/abstracts/1386.html. doi:10.21437/Interspeech.2019-1386.

[30] G. 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. doi:10.21437/Interspeech.2019-1726.

[31] D. Elsner, S. Langer, F. Ritz, R. Mueller, S. Illium, Deep Neural Baselines for Computational Paralinguistics, in: Interspeech 2019, 2019.

[32] P. Wu, S. Rallabandi, A. W. Black, E. Nyberg, Ordinal Triplet Loss: Investigating Sleepiness Detection from Speech, in: Interspeech 2019, 2019, pp. 2403–2407. URL: http://www.isca-speech.org/archive/Interspeech_2019/abstracts/2278.html. doi:10.21437/Interspeech.2019-2278.

[33] S.-L. Yeh, G.-Y. Chao, B.-H. Su, Y.-L. Huang, M.-H. Lin, Y.-C. Tsai, Y.-W. Tai, Z.-C. Lu, C.-Y. Chen, T.-M. Tai, C.-W. Tseng, C.-K. Lee, C.-C. Lee, Using Attention Networks and Adversarial Augmentation for Styrian Dialect Continuous Sleepiness and Baby Sound Recognition, in: Interspeech 2019, 2019, pp. 2398 2402. URL: http://www.isca-speech.org/archive/Interspeech_2019/abstracts/2110.html. doi:10.21437/Interspeech.2019-2110.

[34] J. Fritsch, S. P. Dubagunta, M. Magimai.-Doss, Estimating the Degree of Sleepiness by Integrating Articulatory Feature Knowledge in Raw Waveform Based CNNS, in: ICASSP 2020, Barcelona, Spain, 2020, pp. 6534–6538. URL: https://ieeexplore.ieee.org/document/9053351/. doi:10.1109/ICASSP40776.2020.9053351.

[35] S. Amiriparian, P. Winokurow, V. Karas, S. Ottl, M. Gerczuk, B. Schuller, Unsupervised Representation Learning with Attention and Sequence to Sequence Autoencoders to Predict Sleepiness From Speech, in: MM '20: The 28th ACM International Conference on Multimedia, 2020, pp. 11 17. doi:10.1145/3423327.3423670.

[36] J. V. Egas-Lopez, G. Gosztolya, Deep Neural Network Embeddings for the Estimation of the Degree of Sleepiness, in: ICASSP 2021, Toronto, ON, Canada, 2021, pp. 7288 7292. doi:10.1109/ICASSP39728.2021.9413589.

[37] J. V. Egas-López, R. Busa-Fekete, G. Gosztolya, On the Use of Ensemble X-Vector Embeddings for Improved Sleepiness Detection, in: S. R. M. Prasanna, A. Karpov, K. Samudravijaya, S. S. Agrawal (Eds.), Speech and Computer, Lecture Notes in Computer Science, Springer International Publishing, Cham, 2022, pp. 178–187. doi:10.1007/978-3-031-20980-2_16.

[38] E. L. Campbell, L. Docio-Fernandez, C. Garcia-mateo, A. Wittenborn, J. Krajewski, N. Cummins, Automatic detection of short-term sleepiness state. Sequence-to-Sequence modelling with global attention mechanism., in: Workshop on Speech, Music and Mind, 2022.

[39] E. Hoddes, V. Zarcone, H. Smythe, R. Phillips, W. C. Dement, Quantification of Sleepiness: A New Approach, Psychophysiology 10 (1973) 431 436. URL: http://doi.wiley.com/10.1111/j.1469-8986.1973.tb00801.x. doi:10.1111/j.1469-8986.1973.tb00801.x.

[40] B. Tran, Y. Zhu, X. Liang, J. W. Schwoebel, L. A. Warrenburg, Speech Tasks Relevant to Sleepiness Determined With Deep Transfer Learning, in: ICASSP 2022, 2022, pp. 6937–6941. doi:10.1109/ICASSP43922.2022.9747000.

[41] V. P. Martin, J.-L. Rouas, J.-A. Micoulaud-Franchi, P. Philip, The Objective and Subjective Sleepiness Voice Corpora, in: LREC 2020, Marseille, France, 2020, p. 6525-6533. URL: https://aclanthology.org/2020.lrec-1.803.

[42] V. P. Martin, J.-L. Rouas, F. Boyer, P. Philip, Automatic Speech Recognition systems errors for objective sleepiness detection through voice, in: Interspeech 2021, Brno, 2021, pp. 2476 2480. doi:10.21437/Interspeech.2021-291.

[43] V. P. Martin, J.-L. Rouas, F. Boyer, P. Philip, Automatic Speech Recognition system errors for accident-prone sleepiness detection through voice, in: EUSIPCO 2021, Dublin, 2021, pp. 541–545. doi:10.23919/EUSIPCO54536.2021.9616299.

[44] V. P. Martin, R. Lopez, Y. Dauvilliers, J.-L. Rouas, P. Philip, J.-A. Micoulaud-Franchi, Sleepiness in adults: An umbrella review of a complex construct, Sleep Medicine Reviews 67 (2023) 101718. URL: https://linkinghub.elsevier.com/retrieve/pii/S1087079222001319. doi:10.1016/j.smrv.2022.101718.

[45] V. P. Martin, J. Taillard, J. Rubenstein, P. Philip, R. Lopez, J.-A. Micoulaud-Franchi, Que nous disent les outils de mesure sur la somnolence et l'hypersomnolence chez l'adulte ? Approches historiques et perspectives futures, Médecine du Sommeil 19 (2022) 221–240. URL: https://linkinghub.elsevier.com/retrieve/pii/S1769449322002709. doi:10.1016/j.msom.2022.10.003.

[46] A. W. Maclean, G. C. Fekken, P. Saskin, J. B. Knowles, Psychometric evaluation of the Stanford Sleepiness Scale, Journal of Sleep Research 1 (1992) 35–39. URL: https://onlinelibrary.wiley.com/doi/10.1111/j.1365-2869.1992.tb00006.x. doi:10.1111/j.1365-2869.1992.tb00006.x.

[47] C. Gauld, R. Lopez, P. A. Geoffroy, C. M. Morin, K. Guichard, E. Giroux, Y. Dauvilliers, G. Dumas, P. Philip, J.-A. Micoulaud-Franchi, A systematic analysis of ICSD-3 diagnostic criteria and proposal for further structured iteration, Sleep Medicine Reviews 58 (2021) 101439. URL: https://www.sciencedirect.com/science/article/pii/S1087079221000241. doi:10.1016/j.smrv.2021.101439.

[48] M. Huckvale, A. Beke, M. Ikushima, Prediction of Sleepiness Ratings from Voice by Man and Machine, in: Interspeech 2020, 2020. doi:10.21437/Interspeech.2020-1601.

[49] V. P. Martin, A. Ferron, J.-L. Rouas, P. Philip, "Prediction of Sleepiness Ratings from Voice by Man and Machine": the Endymion replication perceptual study, in: accepted for ICASSP 2023, 2023.

[50] G. R. Norman, The epistemology of clinical reasoning: perspectives from philosophy, psychology, and neuroscience, Academic Medicine 75 (2000) S127 S133. doi:10.1097/00001888-200010001-00041.

[51] V. P. Martin, B. Arnaud, J.-L. Rouas, P. Philip, Does sleepiness influence reading pauses in hypersomniac patients?, in: Speech Prosody 2022, ISCA, 2022, pp. 62 66. doi:10.21437/SpeechProsody.2022-13.

[52] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206 215. URL: http://www.nature.com/articles/s42256-019-0048-x. doi:10.1038/s42256-019-0048-x.

[53] S. Reddy, Explainability and artificial intelligence in medicine, The Lancet Digital Health 4 (2022) e214–e215. URL: https://www.thelancet.com/journals/landig/article/PIIS2589-7500(22)00029-2/fulltext. doi:10.1016/S2589-7500(22)00029-2.

[54] M. W. Johns, A New Method for Measuring Daytime Sleepiness: The Epworth Sleepiness Scale, Sleep 14 (1991) 540–545. doi:https://doi.org/10.1093/sleep/14.6.540.

[55] J.-L. Rouas, T. Shochi, M. Guerry, A. Rilliard, Categorisation of spoken social affects in Japanese: human vs. machine, in: ICPhS, 2019.

[56] K. Sjölander, The Snack Sound Toolkit, Technical Report, 2004. URL: http://www.speech.kth.se/snack/.

[57] V. P. Martin, G. Chapouthier, M. Rieant, J.-L. Rouas, P. Philip, Using reading mistakes as features for sleepiness detection in speech, in: Speech Prosody 2020, Tokyo, Japan, 2020, pp. 985 989. doi:10.21437/SpeechProsody.2020-201.

[58] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi, J. Shi, S. Watanabe, K. Wei, W. Zhang, Y. Zhang, Recent Developments on Espnet Toolkit Boosted By Conformer, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, Toronto, ON, Canada, 2021, pp. 5874–5878. URL:

https://ieeexplore.ieee.org/document/9414858/. doi:10.1109/ICASSP39728.2021.9414858.
[59] S. Galliano, G. Gravier, L. Chaubard, The ESTER 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts, in: Interspeech 2009, 2009, pp. 2583–2586.
[60] V. P. Martin, J.-L. Rouas, A. Basse, B. Caudron, M. Huillet, P. Philip, Est-il possible d'annoter la naturalité des pauses lors de la lecture d'un texte à haute voix ?, in: JEP 2022, 2022. doi:10.21437/JEP.2022-74.
[61] A. S. Zigmond, R. P. Snaith, The hospital anxiety and depression scale, Acta Psychiatrica Scandinavica 67 (1983) 361 370. doi:10.1111/j.1600-0447.1983.tb09716.x.
[62] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research 12 (2011) 2825 2830.
[63] A. Batliner, S. Hantke, B. W. Schuller, Ethics and Good Practice in Computational Paralinguistics, IEEE Transactions on Affective Computing (2020) 1 1. URL: https://ieeexplore.ieee.org/document/9183997/. doi:10.1109/TAFFC.2020.3021015.
[64] R. A. Poldrack, G. Huckins, G. Varoquaux, Establishment of Best Practices for Evidence for Prediction: A Review, JAMA psychiatry 77 (2020) 534–540. doi:10.1001/jamapsychiatry.2019.3671.
[65] J. Horne, C. Burley, We know when we are sleepy: Subjective versus objective measurements of moderate sleepiness in healthy adults, Biological Psychology 83 (2010) 266–268. doi:10.1016/j.biopsycho.2009.12.011.
[66] L. Olson, M. Cole, A. Ambrogetti, Correlations among Epworth Sleepiness Scale scores, multiple sleep latency tests and psychological symptoms, Journal of Sleep Research 7 (1998) 248 253. URL: http://doi.wiley.com/10.1046/j.1365-2869.1998.00123.x. doi:10.1046/j.1365-2869.1998.00123.x.
[67] V. P. Martin, J.-L. Rouas, P. Philip, Détection de la somnolence dans la voix : nouveaux marqueurs et nouvelles stratégies, Traitement Automatique des Langues 61 (2020) 67–90.
[68] V. P. Martin, A. Ferron, J.-L. Rouas, T. Shochi, L. Dupuy, P. Philip, Physiological vs. Subjective sleepiness: what can human hearing estimate better?, in: ICPhS 2023, 2023.
[69] J. Alzubi, A. Nayyar, A. Kumar, Machine Learning from Theory to Algorithms: An Overview, Journal of Physics: Conference Series 1142 (2018) 012012. URL: https://iopscience.iop.org/article/10.1088/1742-6596/1142/1/012012. doi:10.1088/1742-6596/1142/1/012012.
[70] K. Chatzilygeroudis, I. Hatzilygeroudis, I. Perikos, Machine Learning Basics, in: P. Eslambolchilar, A. Komninos, M. Dunlop (Eds.), Intelligent Computing for Interactive System Design, 1 ed., ACM, New York, NY, USA, 2021, pp. 143 193. URL: https://dl.acm.org/doi/10.1145/3447404.3447414. doi:10.1145/3447404.3447414.
[71] D. Sarkar, R. Bali, T. Sharma, Machine Learning Basics, in: Practical Machine Learning with Python, Apress, Berkeley, CA, 2018, pp. 3–65. URL: http://link.springer.com/10.1007/978-1-4842-3207-1_1. doi:10.1007/978-1-4842-3207-1_1.
[72] G. Rebala, A. Ravi, S. Churiwala, Machine Learning Definition and Basics, in: An Introduction to Machine Learning, Springer International Publishing, Cham, 2019, pp. 1–17. URL: http://link.springer.com/10.1007/978-3-030-15729-6_1. doi:10.1007/978-3-030-15729-6_1.
[73] A. Jung, Machine learning: the basics, 2022. URL: https://doi.org/10.1007/978-981-16-8193-6.
[74] E. Stern, J.-A. Micoulaud Franchi, G. Dumas, J. Moreira, S. Mouchabac, J. Maruani, P. Philip, M. Lejoyeux, P. A. Geoffroy, How Can Digital Mental Health Enhance Psychiatry?, The Neuroscientist (2022) 107385842210986. URL: http://journals.sagepub.com/doi/10.1177/10738584221098603. doi:10.1177/10738584221098603.
[75] C. Beaumard, V. P. Martin, Y. Wu, J.-L. Rouas, P. Philip, Automatic detection of schwa in French hypersomniac patients, in: Journée Santé et Intelligence Artificielle (Evènement affilié à PFIA 2023), 2023.

Appendix A. Features example

Two example of feature value averaged across the sessions are reported in Table A.4

Features	Speaker A	Speaker B
Average across sessions	MSL = 1 min ESS = 24 KSS = 6.2	MSL = 14.4 min ESS = 9 KSS = 1.0
Energy slope	4.25e-3	5.84e-3
F.bw2	404,92	343,47
F.bw3	382,98	382,19
F.bw4	378,01	373,76
Ratio +	0,887	0,932
Ratio -	0,113	0,0678

Table A.4: Features implied in the classification of symptoms extracted for two speakers with different sleepiness levels

Appendix B. Decorrelation algorithm example

Let's take the average fundamental frequency F0. This feature is significantly correlated to age, sex, BMI, neck circumference, educational level, and depression levels (Pearson's $r, p < 0.05$). The correlation coefficient and corresponding p-values are reported on the first line of the table below.

This, we estimate $\hat{F}_0$(ge, sex, BMI, neck circ., edu. level, dep.) as a multivariate linear regression, whose coefficients are given in the second table below.

$\hat{F}_0$(sleepiness) is then estimated by par-:

$$\hat{F}_0(\text{sleepiness}) = F_0(\text{measure}) - \hat{F}_0(\text{age, sex, BMI, neck circ., edu. level, dep.})$$

The correlation coefficient between $\hat{F}_0$(sleepiness) and the cofactors confirms the decorrelation of this descriptor with these cofactors (third line of the table below).

	Sex	Age	BMI	Neck circ.	Edu. level	Dep.	Anxiety
r_{before}	-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	-
r_{after}	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)

Appendix C. Performances depending on model hyperparameters

Performances depending on features fusion are reported in Table C.5; performances depending on feature combination on Table C.6; finally, the effect of all the other model parameters are reported in Table C.7.

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 C.5: Max. UAR (%) depending on features fusion

Acoustic	Yes			
ASR	Yes		No	
Pauses	Yes	No	Yes	No
S1	**81.5**	64.4	72.1	68.7
S2	64.8	63.5	71.3	**76.6**
S3	73.2	**79.0**	68.6	**79.0**
SP	70.8	73.4	74.0	**75.3**
SS	66.0	65.5	64.2	**66.2**

Table C.6: Max. UAR (%) depending on features combination

Appendix D. Contribution coefficients of features implied in the classification modeles

The contribution coefficients of features implied in the classification of S1, S2, S3, SP and SS are respectively reported in Tables D.8, D.9, D.10, D.11, and D.12.

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	Tree
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 C.7: Maximum UAR (%, external loop of nested CV) depending on the different elements of the classification model

Category	Name	Median	Avg.	Std.	Min	Max
Acoustic	Energy slope	-0.044	-0.080	0.292	-0.483	0.483
	Formant bw4	0.003	0.193	0.234	-0.038	0.487
Pauses	Ratio +	-0.471	-0.279	0.235	-0.484	0.053
	Ratio -	0.471	0.279	0.235	-0.053	0.484

Table D.8: Median, average, standard deviation, minimum and maximum contribution coefficient of each feature implied in the classification of the objective severe daytime sleep propensity (S1) across the external loop of cross-validation

Category	Name	Median	Avg.	Std.	Min	Max
Acoustic	Energy slope	0.328	0.379	0.432	0.199	0.747
	Formant bw2	0.191	0.144	0.128	-0.131	0.323
	Formant bw3	-0.268	-0.248	0.100	-0.528	-0.024
	Formant bw4	-0.202	-0.192	0.070	-0.304	0.087

Table D.9: Median, average, standard deviation, minimum and maximum contribution coefficient of each feature implied in the classification of subjective perception of severe daytime sleep propensity (S2) across the external loop of cross-validation

Category	Name	Median	Avg.	Std.	Min	Max
Acoustic	Energy slope	-0.5	-0.500	0.000	-0.5	-0.5
	Formant bw4	0.5	0.104	0.491	-0.5	0.5

Table D.10: Median, average, standard deviation, minimum and maximum contribution coefficient of each feature implied in the classification of the average subjective perception of instantaneous sleepiness across the day (S3) across the external loop of cross-validation

Category	Name	Median	Avg.	Std.	Min	Max
Acoustic	Energy slope	-0.173	-0.064	0.190	-0.277	0.309
	Formant bw4	-0.339	-0.350	0.144	-0.640	0.115
	A1	0.012	0.018	0.021	-0.006	0.098
	A2	0.001	0.010	0.023	-0.017	0.095
	H2	-0.026	-0.011	0.029	-0.038	0.081
	H1H2	-0.148	-0.080	0.134	-0.368	0.252
	hnr1	0.006	0.010	0.022	-0.005	0.098
	hnr2	0.019	0.020	0.021	-0.017	0.108
	hnr3	0.019	0.017	0.023	-0.023	0.109
	hnr4	0.026	0.026	0.021	-0.026	0.112
	cpp	-0.158	-0.169	0.099	-0.338	0.277

Table D.11: Median, average, standard deviation, minimum and maximum contribution coefficient of each feature implied in the classification of sleep propensity (SP) across the external loop of cross-validation

Category	Name	Median	Avg.	Std.	Min	Max
Acoustic	F0var	-0.010	-0.010	0.001	-0.018	-0.003
	Energy mean	0.014	0.014	0.002	0.008	0.025
	Energy slope	-0.007	-0.007	0.001	-0.016	-0.004
	Formant fq4	0.004	0.004	0.001	-0.002	0.007
	Formant bw2	-0.018	-0.018	0.001	-0.022	-0.014
	Formant bw3	-0.006	-0.006	0.001	-0.010	-0.002
	Formant bw4	-0.007	-0.007	0.001	-0.011	-0.001
	A1	-0.005	-0.005	0.003	-0.028	0.007
	A2	0.067	0.067	0.005	0.050	0.099
	A3	-0.073	-0.073	0.006	-0.110	-0.053
	H1	-0.044	-0.044	0.003	-0.053	-0.028
	H2	0.017	0.017	0.003	0.004	0.030
	H4	0.016	0.016	0.004	-0.010	0.027
	H1H2	0.016	0.016	0.002	0.010	0.023
	H2H4	0.030	0.030	0.002	0.019	0.039
	hnr1	-0.015	-0.015	0.004	-0.043	-0.004
	hnr2	-0.161	-0.161	0.009	-0.189	-0.117
	hnr3	0.329	0.329	0.006	0.305	0.354
	hnr4	-0.154	-0.154	0.011	-0.196	-0.119
	CPP	-0.004	-0.004	0.001	-0.014	0.001

Table D.12: Median, average, standard deviation, minimum and maximum contribution coefficient of each feature implied in the classification of subjective sleepiness (SS) across the external loop of cross-validation