

Sleep deprivation detected by voice analysis

Etienne Thoret, Thomas Andrillon, Caroline Gauriau, Damien Leger, Daniel

Pressnitzer

▶ To cite this version:

Etienne Thoret, Thomas Andrillon, Caroline Gauriau, Damien Leger, Daniel Pressnitzer. Sleep deprivation detected by voice analysis. 2023. hal-03861009v2

HAL Id: hal-03861009 https://hal.science/hal-03861009v2

Preprint submitted on 10 Oct 2023 (v2), last revised 26 Feb 2024 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers. L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

1	Title
2	 Title: Sleep deprivation detected by voice analysis
3	 Short title: Sleep deprivation detected by voice analysis
4	
5	Authors
6	Etienne Thoret ^{1,4,5} , Thomas Andrillon ^{2,3} , Caroline Gauriau ² , Damien Léger ^{2,†} , Daniel Pressnitzer ^{1,}
7	t
8	
9	Affiliations
10	¹ Laboratoire des systèmes perceptifs, Département d'études cognitives, École normale supérieure, PSL
11	University, CNRS, 75005 Paris, France.
12	
13	² APHP, Hôtel Dieu, Centre du Sommeil et de la Vigilance, 75004 Paris, France & Université Paris Cité,
14	VIFASOM, ERC 7330 Vigilance Fatigue Sommeil et Santé Publique, 75006 PARIS, France.
15	
16	³ Monash Centre for Consciousness & Contemplative Studies, Monash University, Melbourne 3168,
17	Australia.
18	
19	⁴ Aix-Marseille University, CNRS, UMR7061 Perception Representation Image Sound Music (PRISM),
20	UMR7020 Laboratoire d'Informatique et Systèmes (LIS), UMR7089 Institut de Neurosciences de la Timone
21	(INT), Marseille, 13009, France.
22	
23	⁵ Institute of Language Communication and the Brain, Aix-Marseille University, Marseille, France.
24	
25	† Joint last authors.
26	
27	Corresponding author: <u>etiennethoret@gmail.com</u>

28 Abstract (248 words)

Sleep deprivation has an ever-increasing impact on individuals and societies. Yet, to date, there 29 is no quick and objective test for sleep deprivation. Here, we used automated acoustic analyses 30 of the voice to detect sleep deprivation. Building on current machine-learning approaches, we 31 focused on interpretability by introducing two novel ideas: the use of a fully generic auditory 32 representation as input feature space, combined with an interpretation technique based on 33 reverse correlation. The auditory representation consisted of a spectro-temporal modulation 34 analysis derived from neurophysiology. The interpretation method aimed to reveal the regions 35 of the auditory representation that supported the classifiers' decisions. Results showed that 36 generic auditory features could be used to detect sleep deprivation successfully, with an 37 accuracy comparable to state-of-the-art speech features. Furthermore, the interpretation 38 revealed two distinct effects of sleep deprivation on the voice: a change in prosody and a change 39 in timbre. Importantly, the relative balance of the two effects varied widely across individuals, 40 even though the amount of sleep deprivation was controlled, thus confirming the need to 41 characterize sleep deprivation at the individual level. Moreover, while the prosody factor 42 correlated with subjective sleepiness reports, the timbre factor did not, consistent with the 43 presence of both explicit and implicit consequences of sleep deprivation. Overall, the findings 44 show that individual effects of sleep deprivation may be observed in vocal biomarkers. Future 45 investigations correlating such markers with objective physiological measures of sleep 46 deprivation could enable "sleep stethoscopes" for the cost-effective diagnosis of the individual 47 effects of sleep deprivation. 48

50 Author summary (125 words)

Sleep deprivation has an ever-increasing impact on individuals and societies, from accidents to 51 chronic conditions costing billions to health systems. Yet, to date, there is no quick and objective 52 test for sleep deprivation. We show that sleep deprivation can be detected at the individual level 53 with voice recordings. Importantly, we focused on interpretability, which allowed us to identify 54 two independent effects of sleep deprivation on the voice: a change in prosody and a change in 55 timbre. The results also revealed a striking variability in individual reactions to the same 56 deprivation, further confirming the need to consider the effects of sleep deprivation at the 57 individual level. Vocal markers could be correlated to specific underlying physiological factors in 58 future studies, outlining possible cost-effective and non-invasive "sleep stethoscopes". 59

61 Introduction

In the last decade or so, insufficient sleep has become a prominent public health issue, 62 with one third of the adult population sleeping less than six hours per night (1–3). This chronic 63 sleep debt is associated with an increased risk of chronic disease, such as obesity, type 2 64 diabetes, cardiovascular diseases, inflammation, addictions, accidents and cancer (4-8). Sleep 65 debt also increase the risk of developing multiple comorbidities (9). Moreover, more than one 66 worker out of five operates at night and suffers from a high level of sleep deprivation (10), which 67 causes accidents in the workplace or when driving (11). In the present study, we use automated 68 acoustic analyses of the voice to detect sleep deprivation. The aim is not to improve the accuracy 69 of current machine-learning approaches (12–14), but, rather, to build on them to introduce a new 70 focus on interpretability. Ideally, our method should not only detect whether an individual is sleep 71 deprived or not, but also help to formulate specific hypotheses as to the physiological 72 consequences of sleep deprivation for a given individual at a given moment in time. 73

Currently, there are several techniques aiming to measure sleep deprivation and its 74 associated physiological consequences. First, sleep deprivation may be simply assessed in 75 terms of the loss of sleep time, as measured in hours. Remarkably, however, the impact of a 76 given amount of sleep deprivation varies massively across individuals. In laboratory settings 77 where the amount of deprivation could be precisely controlled, up to 90% of the variance in 78 79 cognitive performance was related to individual traits and not to the actual time spent asleep (15, 16). Second, sleep deprivation may also be measured through subjective sleepiness, which 80 participants can explicitly report using rating scales (17–19). However, subjective sleepiness 81 82 could be influenced by other factors than sleep deprivation, such as the time of the day, motivation, or stress. Besides, it is not clear whether reported subjective sleepiness captures 83 the full physiological impact of sleep deprivation, given the variety of the potentially implicit 84 85 processes involved (20). Third, objective methods have been developed to measure tangible

consequences of sleep deprivation. The multiple sleep latency test (21), the gold standard in 86 clinical settings, uses electro-encephalography (EEG) to estimate sleep latency (e.g. the amount 87 of time to go from wake to sleep) along five successive naps sampled every two hours during 88 daytime. The psychomotor vigilance test (22), often used in research settings, tests for the ability 89 to respond quickly to infrequent stimuli, with slower reaction times assumed to be markers of 90 attentional lapses. More recently, new approaches have attempted to measure the concentration 91 of key molecules in the urine, saliva or breath (23). Although these objective methods are 92 complementary to subjective reports, they are often costly, time consuming, or difficult to deploy 93 outside of laboratories. So, whereas there are cheap and fast objective diagnosis tools for other 94 causes of temporary cognitive impairment, such as alcohol or drug abuse, there is currently no 95 96 established means to estimate sleep deprivation effects, at the individual level, in real-life settings. 97

If sleep deprivation could be detected through voice recordings, this would fill this gap by 98 providing a quick, non-invasive, and cost-effective objective measure of sleep deprivation. 99 Indeed, because the voice is easy to record with off-the-shelf equipment, there is a growing 100 interest in finding vocal biomarkers to diagnose a variety of medical conditions (24, 25). For 101 sleep, the idea was first explored by Morris et al. (26). Free speech was produced by sleep 102 deprived participants and rated by the authors. A slowing down of speech and a "flatness of the 103 104 voice" were noted after deprivation. These observations were extended by Harrison and Horne (27), who found that raters blind to the amount of deprivation of the speakers could detect effects 105 106 on the intonation of speech after deprivation. More recently, an experiment using a larger 107 database found that, indeed, raters could detect sleepy versus non sleepy voices with an accuracy above 90% (28). So, it does seem that there are acoustic cues in the voice that reflect 108 109 sleep deprivation and/or sleepiness.

Machine learning has been applied to automate the detection of sleep deprivation and/or 110 sleepiness from the voice. In an early study (29), sleep deprivation was inferred with high 111 accuracy from vocal recordings (86%) but it should be noted that the deprivation was extreme, 112 consisting of 60 hours without sleep, with unknown applicability to the much more common 113 situation of mild sleep deprivation. Two "computational paralinguistic challenges" have since 114 been launched, with sub-challenges aimed at assessing sleepiness from vocal recordings (30, 115 31). We will not review all of the entries to these challenges here, as they are guite technical in 116 nature. To summarize, all of them used a similar framework: i) selection of a set of acoustic 117 features, such as pitch, spectral and cepstral coefficients, duration estimates, and functionals of 118 those features; ii) dimensionality reduction of the feature set; iii) supervised learning of target 119 120 classification using various machine learning techniques, such as support vector machines or neural networks. The best results varied depending on the challenge. Subjective sleepiness 121 proved difficult to predict (28), but the binary categorization of sleepy versus non-sleepy voices 122 could be achieved with high accuracy (over 80%) in the best performing classifiers (32). 123

The framework described above will be familiar -and effective- for many machine learning 124 problems, but it has two major limitations from a neuroscientific perspective. First, the initial 125 selection of features is based on a somewhat arbitrary choice. Often, the choice of features was 126 guided by the "voice flatness" hypothesis (26, 27). However, other, perhaps more subtle acoustic 127 128 markers of sleep deprivation or sleepiness may have been overlooked by human raters. Second, the acoustic features discovered by the classifiers are not necessarily interpretable and can be 129 difficult to relate to plausible mediating mechanisms (14). Interestingly, the best-performing 130 131 system so far used a carefully hand-crafted small feature set inspired from auditory processing 132 models, suggesting that "perceptual" features may be a promising route for sleepiness detection in the voice (32). A more recent study has again attempted to focus on "simple" acoustic 133 134 descriptors for one of the databases of the paralinguistic challenge, with the explicit aim to

facilitate interpretation (33). Accurate classification was possible with the simpler feature set of
 about 20 features, with a resulting accuracy of 76%.

Here, we aim to extend these findings in several ways. First, we use our own vocal 137 database, which has been collected in a controlled laboratory setting where the amount of sleep 138 deprivation could be precisely controlled. Vocal recordings were obtained from reading out loud 139 the same texts for all participants, in random order across participants. This is important to avoid 140 biases confounding sleep deprivation with *e.g.* participant identity, which is easily picked up by 141 classifiers (28, 34). Second, we use a fully generic acoustic feature set, derived from an 142 established model of auditory processing (35). Our audio input representation is based on so-143 called spectro-temporal modulations (STMs). Sounds are first split into separate frequency 144 145 bands, to simulating peripheral auditory filtering, and joint modulations over time and frequency are then estimated, to simulate cortical neural receptive fields (see Methods for further details). 146 While the STM representation was initially motivated by neurophysiological results, it has been 147 successfully applied to various machine-learning problems such as musical instruments 148 classification (36), timbre perception (36, 37), or speech detection and enhancement (38). Third, 149 we apply our own technique to interpret the cues discovered by the classifier (39). This 150 technique, similar in spirit to the reverse correlation method used in neuroscience and 151 psychophysics, identifies the parts of the input representation that have the most weight in the 152 153 classifiers' decisions. The main outcome of the analysis is thus the parts of auditory feature space impacted by sleep deprivation. Fourth, by fitting classifiers to individual participants, we 154 aim to uncover the physiological factors underlying the large and as of yet unexplained variability 155 156 observed in the responses to mild sleep deprivation in normal healthy adults.

- 158 159
- Page 7 of 40

160 **Results**

Twenty-two healthy women between 30-50 years of age (42.7 ± 6.8) were sleep deprived 161 during a controlled laboratory protocol. An all-female experimental group was chosen because 162 the current experiment took place in parallel with a dermatology study (40), but also because 163 such a choice was expected to homogenize the vocal pitch range across participants. After a 164 first "Control night" spent in the laboratory, participants were restricted to no more than 3 hours 165 of sleep per night during two subsequent "Restriction nights", also monitored in the laboratory. 166 Such a sleep restriction is both more ecological than total sleep deprivation and better controlled 167 than observational paradigms. Vocal recordings were obtained throughout the protocol, during 168 reading sessions sampled at different times of the day. These reading sessions occurred either: 169 170 i) right after the control night (no sleep deprivation); or ii) right after the second restriction night (see Methods for details). All participants read 10 minutes of different chapters of the same 171 French classic book: "Le Comte de Monte Christo" (Alexandre Dumas, 1844).. The order of the 172 excerpts was randomized across sessions for each participant to avoid a confound with 173 deprivation. In total, our database consists of 22 healthy participants producing about half an 174 hour of vocal recordings (M=31min, SD=5min) evenly split between before and after two nights 175 of mild sleep deprivation. 176

177

178 Subjective sleepiness reports are variable

179 Sleepiness was self-reported by participants at different times during the day (see 180 Methods) using the Stanford Sleepiness Scale (SSS) questionnaire (19). Figure 1A shows the 181 distributions of SSS ratings. On average, sleep deprivation had an effect on self-reported 182 sleepiness: sleepiness was low right after the control night, but increased after the deprivation 183 nights. This was confirmed by an ANOVA on the SSS, with factors Day (2 levels, before and 184 after deprivation) and Time of Report (4 levels). Both factors had a significant effect, but with a much larger effect size for Day (F(1,46) = 52.14, p<0.001, $\eta_p^2 = 0.221$) compared to Time of report (F(3,92) = 3.07, p=0.029, $\eta_p^2 = 0.048$). Moreover, there was no interaction between Day and Time of report (F(3,92) = 0.59, p=0.621). Because of this lack of interaction, we now consider average SSS values for all Times of Reports in a Day, to focus on the effect of sleep deprivation.

Figure 1B illustrates the data aggregated in that way, with individual changes in 190 sleepiness now identified across the control and sleep deprived day. A remarkable individual 191 variability was obvious in reported sleepiness. Note that this was in spite of our precise control 192 193 of the amount of sleep deprivation, which was equated across all participants. Even so, some 194 participants showed little effect of sleep deprivation, with even cases of decreases in subjective sleepiness after deprivation. Such unexpected effects were observed for all baseline sleepiness, 195 196 low or high, as measured before deprivation. This striking variability is in fact consistent with 197 previous observations involving objective measures of sleep deprivation (16). It also further 198 justifies that vocal biomarkers of sleep deprivation should be investigated at the individual level. 199



200

Fig. 1. A. Subjective sleepiness. Sleepiness was evaluated by self-reports on the Stanford Scale before sleep deprivation (Control) and after two nights of mild sleep deprivation (Sleep

deprived). The abscissa indicates the time of day when sleepiness reports were collected. B.
 Average reported sleepiness before and after sleep restriction. Lines connect data points
 for each participant.

206

207 Acoustic features of speech before and after sleep deprivation are broadly similar

To get a first qualitative overview of the effects of sleep deprivation on the acoustic features of speech, and in particular to test whether deprivation and any obvious average effect on the voice, we first computed STM representations before and after deprivation.

Let us briefly describe the STM representation. At each moment in time, STMs contain 211 the dimensions of frequency, rate, and scale. The frequency dimension, in Hz, reflects the 212 213 spectral content of the sound, similar to a wavelet transform of the temporal waveform. The rate dimension, in Hz, reflects the modulations in sound amplitude in the time domain. Slow 214 modulations have low rates, whereas fast modulations have high rates. Positive rates indicate 215 temporal modulations coupled with downward changes in frequency, whereas negative rates 216 indicate temporal modulations coupled with upward changes in frequency. The scale dimension, 217 in cycle per octave, reflects modulations in the spectral domain. Sounds with fine spectral 218 envelopes have high scale values, while sounds with relatively flat spectral shapes have low 219 scale values. For speech, the dominant rates are between 2 Hz and 8 Hz (41), while dominant 220 scales, related to the harmonic structure of vowels, are around 2 cyc/oct (42). 221



Page 11 of 40

Fig. 2. Acoustic analyses. A. Spectro-Temporal Modulations before sleep deprivation. 224 Projections on the rate-scale and rate-frequency plane are shown. Arbitrary model units. B. As 225 in A., but after sleep deprivation. C. Acoustic difference before and after sleep deprivation, 226 shown as 2 * abs(B-A) / (A+B). Units of percent. **D.** Speech features before (green) and after 227 228 (orange) sleep deprivation. Displayed are four openSMILE features related to average pitch (mean of the fundamental frequency f_0), pitch variation (standard deviation of f_0), voice 229 230 creakiness (Jitter) and voice breathiness (logarithm of the Harmonic to Noise Ratio). Lines 231 connect data points for each participant.

232

223

The full STMs thus have four dimensions of time, frequency, rate, and scale. To have a look at the overall effect of sleep deprivation on acoustic features, we averaged the STMs along the time dimension, separately before and after deprivation. Average STMs before (Fig. 2A) and after (Fig. 2B) deprivation were qualitatively similar. The rate-scale projections showed that, unsurprisingly, high energy in the STMs was focused in regions associated to speech (38). The frequency-rate projection simply showed the average spectrum of our vocal recordings.

To further investigate the candidate acoustic differences caused by deprivation, we 239 subtracted STMs before and after deprivation (Fig. 2C for the population-level results, Fig. S1 240 for individual-level results). At the population level, maximal differences in the rate-scale 241 projection were less than 3%, while differences up to 11% were observed in the frequency-rate 242 projection. At the subject level, differences in the rate-scale projection were around 24.68% on 243 average (SD=6), while differences up to 40.83% on average (SD=12) were observed in the 244 245 frequency-rate projection. Larger differences seem therefore observable at individual level but there is no obvious structure to the differences: they appear noisy and do not necessarily match 246 the STM regions of high energy in speech (see Figure S1). 247

For comparison with the state-of-the art of sleepiness detection from the voice (33), we 249 also computed speech features using the openSMILE library (43). The full 4368 speech features 250 suggested in (33) were extracted (see Methods). Four of them are illustrated in Fig. 2D, 251 averaged before and after deprivation. These features were selected according to the "voice 252 flatness" hypothesis. According to this hypothesis, it could be that sleep deprivation lowered the 253 average pitch of the voice and reduced with its variation. It could also be that the quality of the 254 voice, described with such adjectives as "creakiness" or "breathiness", could systematically 255 change after deprivation. The closest openSMILE correlates of such perceptual descriptors are 256 shown in Figure 2D. Visually, no obvious change was induced by sleep deprivation, with 257 increases or decreases for all four features. 258

At this point, it is unclear whether the raw acoustic differences illustrated in Figure 2 are meaningful compared to the within- and across-participant variability. Also, the choice to illustrate 4 features out of 4368 is somewhat arbitrary. So, it remains to be tested whether the STM or openSMILE features have any predictive power to detect sleep deprivation. To address this point in a principled manner, we now turn to machine-learning, for the new STM representation and also for the openSMILE feature set.

265

266 Detection of sleep deprivation from the voice is possible based on generic auditory 267 features

A first question raised by the present study is whether fully generic auditory features can be used to detect sleep deprivation from the voice. To address this question, the STM representation was used as the input feature space for a standard machine-learning pipeline (13, 14, 36). The dataset was first transformed into train and test splits. We then reduced the high dimensionality of the feature space by means of a principal component analysis (PCA) on the training set (see Methods). The PCA spaces were then fed to a support vector machine classifier (SVM) with a Gaussian kernel (radial basis function). We opted for an SVM and not a
deep-learning architecture mainly because of the relatively modest size of our dataset, but also
because SVMs have been shown to outperform more complex classifiers in similar tasks (14).
The performance of the SVM was evaluated with Balanced Accuracy (BAcc, see Methods).

At the population level, two cross-validation strategies were used. First, a Leave-One-278 Subject-Out (LOSO) strategy, in which one subject was left out of the training set and constituted 279 the test set. The procedure was repeated for each participant. This procedure is the most 280 stringent test of generalization of prediction for unknown participants. However, in small datasets 281 with a large amount of individual variability, it has been argued that LOSO may be inappropriate 282 (44). This is likely the case for our dataset, with 22 participants and a large expected variability 283 for the effects of sleep deprivation. Thus, we also report cross-validation using a 50-times 284 repeated splitting of 25% of the data (50-splits, 25% test) randomly selected among the whole 285 pool of participants, as suggested in (44). At the participant level, the LOSO strategy does not 286 make sense, so only the 50-splits, 25% test validation was applied. 287





feature space. Two cross-validation procedures are reported (see text). Error bars show standard deviations. Stars indicate the significance level of *t*-tests against chance level (** < .01; *** < .001). **B.** Balanced Accuracies for the classifiers tuned to individual participants, obtained with the 50-splits, 25% test cross-validation procedure. Participants are ranked according to classification accuracy.

297

Classification performance is shown in Figure 3. At the population level, the classifier was 298 able to detect sleep deprivation significantly above chance (50-splits, 20% test: BAcc, M=.77, 299 SD=.01, t-test against .5: t(49) = 145.27 p < 0.001; LOSO: BAcc, M=.56, SD=.09, t-test against 300 .5: t(21) = 3.01, p = .006). This seems on par with the state of the art obtained with different 301 302 speech databases (32, 33). Interestingly, and as expected from the sizeable individual variability observed in the SSS reports, the same machine-learning pipeline was more accurate when 303 applied at the individual level (BAcc, M=.86, SD=.09). Noticeably, for half of the participants, the 304 classifiers' accuracies displayed BAccs above .9, outperforming the state of the art and matching 305 human performance on a similar task (28). For two participants, the classifiers' accuracies were 306 relatively poor. Participant #1 displayed a decrease in sleepiness after deprivation (-0.75 for the 307 sleepiness ratings averaged after and before deprivation), and was the only participant to exhibit 308 such a trend in the group for which vocal recordings were available (another participant exhibited 309 310 such a decrease in Figure 1B, but was could not be included in the vocal analysis). This may have contributed to the poor accuracy of the classifier. Participant #2 did exhibit an increase in 311 312 sleepiness after deprivation (+0.75), so there are no obvious reasons for the classifier's poor 313 performance in this case.

Overall, this shows that there is enough information in the fully generic STM representation of vocal recordings to detect mild sleep deprivation in otherwise normal and healthy participants. The classification performance at the population level is poor using a LOSO 317 cross-validation procedure, so the generalizability of our approach across speakers is not 318 warranted. However, performance is generally excellent at the individual level, strengthening the 319 idea that individual variability is key when considering vocal correlates of sleep deprivation.

320

321 Using standard speech features does not improve sleep deprivation detection accuracy

Even if the STM representation successfully supported sleep detection at the individual level, it could be that it missed important speech features such as "pitch" or "pitch variation", which are at the core of the "voice flatness" hypothesis and are part of most automatic sleepiness detection pipelines.

To investigate this possibility, we used the full openSMILE feature set (4368 features) as input feature space. We then applied the same classification pipeline as for the STM representation, consisting of dimensionality reduction followed by a Gaussian kernel SVM. This resulted in a pipeline matching the state-of-the art for sleepiness (33) while allowing comparison between the two input spaces.

Results are displayed in Figure 4. At the population level, the openSMILE classifier 331 accuracy was similar to the STM classifier (50-splits, 20% test: BAcc, M=.70 SD=.01, t-test 332 against .5: t(49)=91.6, p < 0.01; LOSO: BAcc, M=.56, SD=.09, t-test against .5: t(21)=2.98, p = 333 .007). At the individual level, the accuracies of the openSMILE classifiers were on average 334 poorer than those observed with the STM classifiers (BAcc, M=0.67, SD=0.9). The correlation 335 between classification performance using STM or openSMILE feature was low (r(20) = .34), 336 p=.11). Interestingly, however, participants #1 and #2 for whom poor classification performance 337 338 was observed using the STM input feature space also displayed poor classification using the 339 openSMILE input feature space.

These results show that, for our voice database at least, using standard speech features decreased the accuracy of sleep deprivation detection. The relevant information to detect sleep deprivation from the voice was thus better expressed in the STM representation, with the added
 benefit, from our perspective, that generic auditory features should be easier to interpret. We
 thus now focus on the STM representation to interpret the features used for classification.



Fig. 4. Machine learning classification results with openSMILE input features. Format as
in Fig. 3. In particular, for B., participants' labels (#) are identical to Fig. 3.

349

350 Interpreting classifiers to identify vocal biomarkers of sleep deprivation

To gain insight about the nature of the acoustic features distinguishing speech before and 351 after sleep deprivation, we probed the trained classifiers with STM using a recent interpretation 352 technique based on reverse correlation (39). Briefly, the technique consists in randomly 353 perturbing the input to the trained classifier, over thousands of trials, and then averaging all of 354 the noisy representations leading to correct classification. This aims to identify the portion of the 355 input that participates the most to the classifier's performance. The input representation was 356 perturbed using additive noise in the PCA-reduced feature space (45). Averaging all masks 357 leading to a correct classification decision revealed, in our case, the discriminative features of a 358 voice after deprivation compared to before deprivation (for details, see Methods and 39). 359

As a preliminary step, we evaluated the consistency of the interpretation masks. Because 360 of our cross-validation technique, 50 classifiers were fitted either for the whole dataset for the 361 population-level classifier or for each participant's classifier. To check internal consistency, we 362 computed the pairwise Pearson's correlation coefficients between all 50 interpretation maps. At 363 the population-level, this "consistency correlation" was low albeit significantly above chance 364 (r(22527): M = .20, SD=.34; all but 28 over 1225 pairwise correlations were significant, p<.05)365 which is consistent with the large variability suspected across listeners. At the participant-level, 366 however, consistency correlations were very high (r(22527): M=.91, SD=.06, min=.73; all but 3 367 over 26950 pairwise correlations were significant, p<.05). Furthermore, because individual 368 classifiers varied in accuracy, we could check whether the consistency of the interpretation 369 improved with accuracy. As expected, the correlation between BAccs and consistency 370 correlation was strong (r(20)=.71, p=.0003). These consistency results confirm that caution 371 should be applied when considering population-level interpretations, but that individual results 372 are robust and can be interpreted. 373

Figure 5 shows the interpretation maps for the population-level classifier. Maps should 374 be read as follows: red areas correspond to STM features where the presence of energy is 375 associated with sleep deprivation for the classifier, whereas blue areas represent STM features 376 where the absence of energy is associated to sleep deprivation for the classifier. For the 377 population-level map, the rate-scale projection resembles the raw difference before and after 378 deprivation, although less noisy, whereas the frequency-rate projection does not match such 379 raw acoustic differences (compare with Fig. 2C). As these population-level interpretations are 380 381 not robust, we simply show them for illustrative purposes and refrain from further description of their features. 382



Fig. 5. Interpretation of the population-level classifier. Discriminative features (see main text) are shown in the input STM space, for the rate-scale and frequency-scale projections. Red areas indicate features positively associated to sleep deprivation by the classifier. Blue areas correspond to features negatively associated to sleep deprivation by the classifier. Color bar indicate the averaged value of the reverse correlation mask. Values are low because of the relative low consistency of the interpretation masks for this population-level classifier.

391

Figure 6A shows all individual classifiers on the rate-scale projection, ordered along 392 increasing accuracy (BAcc) of the corresponding classifier. We chose to interpret in priority the 393 394 rate-scale projections, as is done in most speech applications (38). The frequency-rate projections are provided as Figure S3. The main feature of the results is the striking diversity of 395 the individual maps, which is not related to classifier accuracy in any obvious manner. For some 396 participants, sleep deprivation was detected through a reduction in energy over a range of STM 397 features (blue color), consistent with a "flattening" of speech modulations. But the opposite was 398 also observed for other participants (red color). Moreover, the details of the discriminative 399 features also varied across participants. As shown before, these details are robust and warrant 400 interpretation. 401



404

Fig. 6. Interpretation of the participant-level classifiers. A. As for Figure 5, but for individual 405 participants identified by their participant #. B. Projection of participants # in the interpretation-406 PCA space of all participant's masks (see text for details). C. and D. Variance of the idealized 407 masks along the first two dimensions of the interpretation-PCA. Idealized masks are obtained 408 by first sampling the PCA latent space between -2 and 2 for the two first dimensions with 30 409 values and then inverting the latent space into the input feature space by using the inverse 410 transform of the PCA. Red areas show the discriminative features that vary the most along each 411 interpretation-PCA dimension. Units: variance in the feature space. 412

To get a better understanding of this variability across individual maps, we performed a 414 PCA on the maps themselves, which we will term interpretation-PCA for clarity. A first 415 interpretation-PCA dimension explained 35.9% of the variance, while a second dimension 416 explained 24.2% of the variance. There is a drop for all other dimensions (N=3) which explain 417 less that 13% of the variance, see Figure S4. Participants ordered on the first two interpretation-418 PCA dimensions are shown in Figure 5B. We computed the variance of all STM features along 419 each interpretation-PCA dimension, to visualize the features that distinguished the interpretation 420 maps along these main axes of variation. Results are shown in Figure 6C, D. The features 421 defining the first interpretation-PCA dimension were clustered between rates of about 2 Hz to 422 8 Hz, which is exactly the amplitude modulation range corresponding to speech prosody and 423 424 syllabic rate (41). This shows that the amplitude modulation characteristics of speech was affected by sleep deprivation. Importantly, depending on the individual, the classifiers used the 425 presence or absence of energy around these rates to detect sleep deprivation. This shows that 426 while some participants spoke in a "flattened" voice after deprivation, consistent with classic 427 hypotheses (26, 33), others instead spoke in a more "animated" voice after deprivation. The 428 features defining the second interpretation-PCA dimension clustered at very low rates and 429 covered a broad scale range, peaking at about 2 cyc/oct. This corresponds to long-term spectral 430 431 characteristics of speech and vowel sounds. In speech, such timbre-like features are determined 432 by the precise shape of the various resonators inside the vocal tract, such as the throat and nasal cavities: by filtering the sound produced by the vocal folds, resonators impose formants 433 that impact the timbre of vowels and other speech sounds. 434

435

436 **Correlation with subjective sleepiness reports**

All participants were subjected to the exact same amount of sleep deprivation.
 Nevertheless, their subjective sleepiness reports varied widely (Fig. 1). We investigated whether

the variability in subjective sleepiness reports could be accounted for by characteristics of theindividual machine-learning classifiers.

First, we simply correlated the individual classifier's accuracies to the individual SSS reports after sleep deprivation. If subjective sleepiness was a full measure of the impact of sleep deprivation, we would expect a high correlation between the classifier's accuracy and SSS reports. Results are shown in Fig. 7A. There was no significant correlation between BAccs and SSS reports (r(20)=.32, p=.14, BF₁₀=.47), suggesting that subjective sleepiness did not express all of the objective effects of sleep deprivation, at least as captured by our voice classifiers.

Next, we investigated whether the classifier's interpretation maps could account for the 447 SSS reports variability. In particular, we reasoned that the prosodic and rhythmic changes 448 449 captured by the first interpretation-PCA dimension could be due to cognitive factors, inducing flattened or animated speech. Such factors could be explicit to participants - if only by self-450 monitoring their own speech rate and intonation. In contrast, the timbre cues captured by the 451 second interpretation-PCA dimension could be more subtle and remain implicit. Results are 452 shown in Fig. 7B & 7C. Consistent with our hypothesis, we observed a moderate but significant 453 correlation between the location of participants on the first interpretation dimension and 454 sleepiness reports (r(20)=-.44, p=.03, BF₁₀=1.34). In contrast, the location of participants on the 455 second interpretation dimension did not show any significant correlation with sleepiness reports 456 457 $(r(20)=.19, p=.38, BF_{10}=.23).$

Finally, to assess the full information contained in the interpretation maps, we fitted a linear model that used coordinates on both interpretation-PCA dimensions to predict SSS scores after deprivation (see Methods). Results showed that it was possible to predict sleepiness from interpretation maps (R²: M=.29, SD=.18) significantly above chance (two-sample t-test to 0: p<.00001). The correlation remained moderate, however, with still a sizeable part of the variance unexplained.





Fig. 7. Relation between subjective sleepiness and voice classifiers. A. Subjective sleepiness is plotted as a function of balanced accuracy of each participant-level classifier. B. Subjective sleepiness is plotted as a function of the coordinate of each participant-level classifier on the first dimension of the interpretation-PCA space. C. As in B., but for the second dimension of the interpretation-PCA space.

471

472 **Discussion**

473 Summary of findings

We ran a sleep deprivation protocol with normal and healthy participants, collecting 474 subjective reports of sleepiness plus vocal recordings before and after deprivation. After two 475 nights of mild sleep deprivation, subjective sleepiness increased on average, although with 476 striking individual differences-including some participants even reporting decreases in 477 subjective sleepiness after deprivation. Nevertheless, sleep deprivation could be detected 478 accurately by means of machine-learning analysis of vocal recordings. Classification was most 479 480 accurate at the individual level, with 85% balanced accuracy on average. Importantly, such a classification was based on a fully generic auditory representation. This allowed us to interpret 481 the discriminative features discovered by classifiers to detect sleep deprivation. Two broad 482 483 classes of features were revealed: changes in temporal modulations within the rhythmic range

characteristic of speech sentences, and changes in spectral modulations within the timbre range
of speech sounds. Furthermore, the interpretation maps could account for some of the variability
in subjective sleepiness reports, which were correlated to the changes in temporal modulations
("flattened" or "animated" voice).

488

489 Candidate mechanisms underlying the vocal biomarkers of sleep deprivation

The individual classifiers using the STM input features learnt to detect sleep deprivation with a high accuracy, matching human performance, based on two classes of auditory features: temporal modulations in the 2 Hz to 8 Hz range, and spectral modulations around 2 cyc/oct. We now speculatively relate these vocal features to two classes of well-established neurophysiological effects of sleep deprivation.

The temporal modulation features associated to sleep deprivation were in a range which 495 has been robustly found as characteristic of speech across a variety of languages, to the extent 496 that they have been described as "universal rhythmic properties of human speech" (41). Such a 497 universal rhythm is imposed by the biomechanical constraints of the vocal apparatus and by the 498 neurodynamics of its control and perception systems. The changes in speech rhythms observed 499 after sleep deprivation could thus result from a temporary impairment of the cognitive control of 500 the speech production process. Sleep deprivation impacts cognitive function (20), presumably 501 502 through changes in glucose consumption in frontal and motor brain regions (46, 47). Accordingly, previous studies showed lower activity in the dorsolateral prefrontal cortex and in 503 the intraparietal sulcus in cognitive tasks requiring attention, with large inter-individual variability 504 505 (48). A reduced connectivity was also observed within the default mode network, the dorsal attention network, and the auditory, visual and motor network following sleep deprivation (47, 506 49, 50). Finally, extended wakefulness has been associated with an increase in the intrusion of 507 508 sleep-like patterns of brain activity in wakefulness (51, 52). All these results suggest that sleep

deprivation is akin to a minor cognitive frontal dysfunction, and may thus plausibly affect the fluency of vocal production. Interestingly, compensatory responses were also observed in cognitive tasks, which may explain why some of our participants responded to deprivation with less speech modulation, consistent with the classic "flattened voice" hypothesis (26, 27), whereas others unexpectedly responded with speech over-modulation and instead produced an "animated voice" after deprivation.

The spectral modulation changes detected by our classifiers were consistent with 515 changes in the timbre of speech sounds, and in particular vowel sounds (35, 38, 42). Such 516 sounds acquire their distinctive spectral envelopes by means of the resonances of the vocal 517 tract, including the throat and nasal cavities. Inflammation of the throat and nose could be 518 519 responsible for these changes in timbre. Sleep deprivation is known to trigger an immune response leading to inflammation. A cortisol increment can be observed after a single night of 520 sleep deprivation (8, 53, 54), so is plausible in our protocol that included two nights of mild sleep 521 deprivation. In terms of mechanisms, sleep restriction and deprivation disturb the normal 522 secretion of hormones like cortisol or testosterone, and is associated with increased rates of 523 interleukin-6 and CRP as assessed on salivary samples in normal subjects. This inflammatory 524 response could be linked to an elevated blood pressure following sleep deprivation (55) and 525 could affect the vocal tract and plausibly impact the spectral envelope of speech. It should be 526 527 noted that other variables, such as changes in hydration or food intake due to deprivation, might also impact characteristics of the vocal apparatus and induce timbre changes instead or in 528 addition to putative inflammation. Such additional variables were not controlled in our protocol 529

530

531 Limitations of the study

532 There are both technical and conceptual limitations to the present study. We chose to 533 use a controlled protocol to precisely equate sleep deprivation in our participants, but this came

at the expense of a relatively small dataset compared to the online databases used by machine-534 learning challenges (30, 31). Our protocol prevented biases in the database, such as associating 535 the identity of speakers with the amount of sleep deprivation (28), but also limited our choice of 536 possible machine-learning techniques to perform the classification. We thus used an SVM 537 classifier, and not potentially more powerful deep-learning architectures. We note however that 538 in the studies that compared SVMs with other classifier types, SVM performed best, including in 539 state-of-the-art studies (14, 32, 33). In any case, the interpretation method we used could be 540 applied to any kind of classifier (39), including more complex ones. 541

All participants were female, mainly for practical reasons. Sleep deprivation might affect 542 females and males differently, in particular with respect to inflammation, although the evidence 543 544 is still mixed (56). The generalizability of our findings to males thus remains to be tested experimentally. In addition, the modest performance observed for population-level classifiers 545 limits the generalization of our approach to unknown speakers, which would be desirable for 546 practical use cases involving pre-trained classifiers. However, this also confirms the interest to 547 apply interpretation techniques at the indivi` dual level, to capture the variability that seems 548 inherent to the effects of speech deprivation. 549

The feature set we used was a generic auditory representation, which is a major 550 difference with previous machine-learning oriented studies. On the one hand, some studies were 551 552 fully data-driven and selected the best-performing features from thousands of speech descriptors. The resulting features were often difficult to interpret. On the other hand, there were 553 also studies using a small set of features, but these features were carefully hand-crafted and 554 555 potentially lacked genericity. Our approach represents a trade-off between these two ideas: we applied a data-driven approach to select a small subset of features, but because these features 556 were from a generic representation, they remained interpretable. A clear limitation is that we did 557 558 not include features related to pitch or functionals of pitch such as contour features, which have been repeatedly shown to be useful for sleepiness detection (14, 32, 57). However, average estimates of pitch and pitch variation (Fig. 2D) suggested that there were no obvious effect on these features in our database. Furthermore, our classification pipeline applied to standard speech features performed worse than using the STM representation. We believe that these omissions were compensated by the richness of the STM representation. Pitch and pitch functionals will in fact be indirectly reflected in the STMs, which analyses sounds over a broad range of temporal scales simultaneously.

The possible physiological mechanisms that we put forward as a mediation between 566 sleep deprivation and vocal features have to be considered as fully speculative for now. We did 567 not collect the objective measures required to confirm or infirm these interpretations. The 568 569 cognitive factor could be assessed with objective behavioral measures, such as the psychomotor vigilance test (22), or with brain imaging data (46, 47). The inflammatory factor 570 could be assessed by biological analyses of e.g. cortisol in the saliva (8, 54). Because we have 571 not gathered such measurements, we can only argue that both minor cognitive dysfunction and 572 inflammation effects are likely for our participants as a group. In any case, the present study is 573 the first one to suggest that such factors may be measured at the individual level from voice 574 biomarkers, and it raises the possibility for future investigations to confirm or reject this 575 hypothesis by actually correlating vocal features with more invasive objective markers. 576

The classification task we investigated consisted only in detecting whether a vocal recording was performed before or after sleep deprivation. We did not attempt to decode the effect of more subtle factors on the voice, such as the time of the day, which would reflect the interactions between circadian rhythms and sleep deprivation. These interactions have been shown in a recent study (58), albeit using a more more severe deprivation protocol (60 hours without sleep). Unfortunately, our experimental design does not provide the statistical power to examine within-day variations before sleep deprivation (half the dataset) or the interaction between within-day variations and deprivation (second-order effect). In the same study, a regression approach was implemented to provide predictions beyond binary classification. Interestingly, this approach was successful only for predicting objective measures, such as sleep latency, but failed for subjective reports. This is consistent with the claim that subjective scales incompletely characterize the full effects of sleep deprivation and can usefully be complemented by objective measures such as voice analysis. In any case, as we did not collect objective measures of sleepiness beyond the voice, we did not attempt a regression analysis.

Finally, on a conceptual level, we wish to raise a basic but inescapable limitation of any 591 study of sleep deprivation. Sleep deprivation may be defined, as we did, by the amount of sleep 592 available to each individual. However, as has been repeatedly pointed out and again observed 593 594 here, there is a remarkable diversity of responses to the same amount of sleep deprivation. Thus, it should not be expected that any one measure will capture all of the effects of sleep 595 deprivation. Subjective reports may capture explicit feelings of fatigue, but be blind to implicit 596 effects (58). With objective measures, which are by necessity indirect, there is an issue with 597 interpreting negative outcomes. In our case for instance, how to interpret a relatively poor 598 accuracy for a sleep deprivation classifier, such as was observed for two participants? It cannot 599 be decided whether this poor accuracy showed that sleep deprivation had no effect on these 600 participants, or that sleep deprivation had effects that were not expressed in the voice, or that 601 the classifiers simply failed for technical reasons. Measuring multiple markers of sleep 602 deprivation, including the novel ones we suggest, and incorporate them into a holistic model of 603 the neurophysiological effects of sleep deprivation seems to be a promising way forward. 604

605

606 **Perspectives**

607 Keeping these limitations in mind, the demonstration of vocal biomarkers for sleep 608 deprivation could have major clinical implications. Subjective sleepiness reports do not capture

the whole effect of a lack of sleep (58). Moreover, such reports rely on the honest cooperation 609 of participants, which is not a given if self-reports of excessive sleepiness can have negative 610 work-related or financial consequences for the individual. Objective correlates of sleepiness 611 exist (21, 22), but vocal biomarkers would represent a considerably cheaper and faster 612 alternative, requiring no specialized equipment and increasing their practicality for real-life 613 clinical assessment. Crucially, our technique also goes beyond the simple binary detection of 614 sleep deprivation: thanks to the application of interpretability techniques (39), we suggest that 615 different neurophysiological processes related to sleep deprivation may be untangled through 616 the voice alone. Such measures could in turn be used to design interventions tailored to each 617 individual and situation, if the effects of sleep deprivation needed to be temporarily alleviated for 618 619 instance. More generally, there is a growing realization that interpretability is key to future clinical applications of artificial intelligence, as both patients and clinicians would understandably want 620 to understand the reason for a diagnostic (59). For application to real-life settings, it is particularly 621 interesting to identify features that do not correlate with subjective sleepiness, as one of the 622 biggest dangers of sleep loss is the partial agnosia for one's own sleepiness. 623

To finish, it is useful to point out that the methodological pipeline we introduced here is fully generic, as the audio features representation used is itself generic and the interpretation method can be applied to any classifier. Therefore, the present study could pave the way for future investigations of vocal biomarkers over the broad range of fundamental or clinical applications that are currently only starting to be considered (24, 25).

629

630 Materials and Methods

631 *Ethics statement*

The study was conducted according to French regulations on human research including
 agreements from the Hotel-Dieu Hospital Ethics Committee (CPP IIe de France 1 - N° 2017-

sept.-13690), with signed consent from participants who received financial compensation. Our
 protocol was conducted in accordance with the 2016 version of the Declaration of Helsinki and
 the ICH guidelines for Good Clinical Practice.

637 **Experimental design**

A group of twenty-four healthy women between 30-50 years old (42.7 \pm 6.8) took part in the experiment. This study was part of a dermatological study and only Caucasian phototypes I-IV (Fitzpatrick classification) were recruited. Participants were non-smokers and did not report a history of substance abuse. They had a Body Mass Index (BMI) between 19 and 25, no sleep disorders or chronic disease, no daytime vigilance issues (Epworth Sleepiness Scale \leq 10), and were not under any medical treatment (exclusion criteria).

Before the experiment, participants wore an actigraph for 7 days and were instructed to 644 maintain a regular sleep-wake behavior with their usual 7-8 h of sleep (i.e., in bed from 23:00-645 01:00 until 07:00-09:00). The compliance with these recommendations was verified through the 646 actigraphic recordings (MW8, CamTech; UK) that were inspected by the research team at the 647 participant's arrival the morning before the first night of sleep restriction (day 1). No sleep 648 episodes were detected outside of the scheduled experimental time in bed (see 40 for details). 649 The protocol lasted for 3 days (day 1: before sleep restriction; day 2: during sleep restriction; 650 day 3: after sleep restriction), which included 2 night of sleep deprivation (at the end of day 1 651 652 and 2). During the "sleep restriction" session, the participants were instructed to restrict their sleep time to 3h for 2 consecutive nights (i.e., in bed from 03:00 to 06:00) and to follow their 653 usual routine outside the laboratory. After the second sleep-restricted night (day 3), the 654 655 participants went to the laboratory on the morning and their actigraphy recordings were immediately analysed to ensure their compliance with the imposed sleep-wake hours. During 656 day 1 (after habitual sleep and before sleep restriction: baseline condition) and day 3 of each 657 658 session, the participants remained in the sleep laboratory from 09:00 to 19:00 under continuous

supervision. In order to help the participants stay awake, from the moment they left the laboratory 659 at the end of day 1 until their return to the laboratory at the beginning of day 3 at 09:00, two 660 investigators exchanged text messages with the participants at random times during the entire 661 period outside of the laboratory. Text messages were sent throughout the night (except during 662 the period where participants were instructed to sleep, that is between 3 and 6 a.m.). Participants 663 had to respond right after receiving these messages. In case of an absence of response, 664 participants were immediately called on their personal phone. For lunch in the laboratory (day 1 665 and 3), participants received controlled meals consisting of a maximum of 2,500 calories/day 666 with a balanced proportion of nutrients (protein, fat, and carbohydrates). 667

668 Voice recordings

During day 1 (before sleep deprivation) and day 3 (after), at three different times during the day (9am, 3pm, 5 pm), participants were seated and instructed to read 10 minutes of different chapters of the same French classic book: "Le Comte de Monte Christo" (Alexandre Dumas, 1844). Their voice was recorded with a portable recorder (Zoom H1/MB, stereo-recording). Then, during one minute, participants produced free speech, but these recordings were not used in the present analyses. Two participants had to be discarded at this stage, as technical issues prevented the completion of all recording sessions.

676 Baseline Speech Feature Set

We computed basic speech features using the openSMILE library (43), in the configuration recommended for the Interspeech 2011 challenge. This feature set has been used for in stateof-the-art studies detecting sleepiness from voice recordings (33). It consists of 59 low-level descriptors, including 4 energy descriptors, 50 spectral descriptors, and 5 voice descriptors. These descriptors were then combined with 33 base functionals and 5 f0 functionals, resulting in a total of 4,368 features.

683 Spectro-Temporal Modulations (STM)

The sound files, initially sampled at 44.1 kHz, were down-sampled to 16 kHz. Spectro-Temporal 684 Modulations (STMs) were computed with our own toolkit which is directly adapted from the 685 standard NSL Toolbox (35). Sounds were processed through a bank of 128 constant-Q 686 asymmetric bandpass filters equally spaced on a logarithmic frequency scale spanning 5.3 687 octaves, which resulted in an auditory spectrogram, a two-dimensional time-frequency array. 688 The STM were then computed by applying a spectro-temporal modulation filterbank to the 689 auditory spectrogram. We generally followed the procedure detailed in (36), with minor 690 adaptations. A 2D Fourier transform was first applied to the spectrogram resulting in a two-691 dimensional array, also called Modulation Power Spectrum (MPS) (60) whose dimensions were 692 spectral modulation (scale) and temporal modulation (rate). Then, the STM representation was 693 694 derived by filtering the MPS according to different rates and scales and then transforming back to the time-frequency domain. We chose the following scale (s) and rate (r) center values as 2D 695 Gaussian filters to generate the STMs: *s* = [0.71, 1.0, 1.41, 2.00, 2.83, 4.00, 5.66, 8.00] cyc/oct, 696 $r = \pm [.25, .5, 1, 2, 4, 5.70, 8, 11.3, 16, 22.6, 32]$ Hz. Such a range covers the relevant spectral 697 and temporal modulations of speech sounds as already used in different studies (61). The 698 resulting representation thus corresponds to a 4D matrix with dimensions of time, frequency, 699 scale, and rate. 700

701 Classification pipeline

For all recordings, STMs were computed and used as the input feature space. The STM feature space was sampled with 22 rates * 8 scales * 128 frequencies per 3841 temporal frames corresponding to epochs of 15 seconds, amounting to 22528 features for every sample. Standard machine-learning pipeline were used (13, 14, 36) to evaluate the ability to predict a whether a voice sample is from the sleep deprived class.

First, the whole dataset was randomly separated into a training set and a testing set, either by randomly holding 25% of the data into the testing set or, only at population level, by

holding-out the data from one subject to define the training and the testing in a Leave-One-709 Subject-Out (LOSO) cross-validation procedure. We then reduced this high dimensionality of 710 the feature space by means of a principal component analysis (PCA). At the population level, 711 we trained a PCA on the whole dataset and retained the 250 main dimensions, explaining 99% 712 of the variance. We further checked that the exact choice of PCA dimensions did not affect our 713 conclusions, about the performance but also about the interpretation of the classifiers (see 714 Figure S2). At the participant level, for each participant we trained a PCA on the data from all 715 other participants, to reduce a possible contamination of the reduced space by peculiarities of 716 the considered participant. We next retained the 30 main dimensions of the PCA. The number 717 of PCA dimensions in this case was chosen empirically, so that the reduced feature space still 718 719 explained more than 90% of the variance and provided a dimensionality lower than the number of samples available for each participant (between 98 and 194 samples of 15 sec. each), to 720 avoid overfitting. We checked that the exact choice of PCA dimensions did not affect our 721 conclusions, in particular on the interpreted features that are consistent for PCA dimensions 722 above 30. 723

The PCA spaces were then fed to a support vector machine classifier (SVM) with a 724 gaussian kernel (radial basis function). The training set was used to fit the SVM through an 725 hyperparameter grid-search, using a stratified 5-folds cross-validation. The fitted SVM was then 726 727 evaluated on the testing set by computing Balanced Accuracy (BAcc, defined as the average of true positive rate, or sensitivity, with true negative rate, or specificity). For the randomly selected 728 train/test split, we repeated the fitting procedure 50 times, generating 50 distinct train/test sets 729 730 for both the population and individual levels (denoted 50-splits, 25% test). In the Leave-One-Subject-Out (LOSO) approach, we replicated the fitting procedure with 22 subjects, each 731 excluded once and designated as the testing set. In each instance, we computed the final 732 733 balanced accuracies and then averaged them across either the 50 different train/test splits (int the case of 50-splits, 25% test) or across the 22 subjects (in the case of LOSO). Lastly, for each cross-validation procedure, we conducted a t-test against the threshold of 0.5 using the distributions of balanced accuracies. This allowed us to evaluate the classifier's capability to predict sleepiness from voice, assessing its performance compared to random chance. All the classification pipelines from PCA to RBF + SVM are implemented with the sci-kit learn library (Pedregosa et al., 2011).

740

741 Interpretation of the classifiers

Each classifier fitted in the study is probed with the reverse correlation technique which 742 provides an interpretation of which features are relevant in the sense of the classifier. 743 744 Theoretically, for each feature of the input space, a correlation is made between the array of random values from each noise sample with the array of decision values, 0 or 1, 0 corresponding 745 to excerpts classified as before sleep restriction and 1 to excerpts recorded after sleep 746 restriction. Here, as the noise were of null average in the feature space, we simply subtracted 747 the average values of the noises that led to a classification in the class 'after restriction' with the 748 average values that led to classification in the class 'before restriction' (62). We refer to our 749 method paper for a full description of the method (39). Here, we used the version of the method 750 which consists of pseudo-random noise as perturbation at the input of the system. Pseudo-751 random noises allow to accurately fool the classifier while using a white noise may implicate 752 complication as the classifier can tend to classify all the stimuli + noise excerpt in only one class. 753 One specificity of this method is that it requires a large number of trials to provide an accurate 754 755 description of the importance of each feature in the input space. Here we chose to use a number of trials equal to 100 times the number of samples which represents between 9800 and 20000 756 trials. Each interpretation provides a "interpretation mask" which are composed of positive and 757

negative values, positive values correspond to features which are diagnostic of sleep loss and
 negative ones conversely.

For each classification task, 50 classifiers were fitted. In order to test the independence of the prediction accuracy from the 50 different random training set. Each of these 50 classifiers were interpreted with the previously described method and a second test was then performed in order to test the similarities between the 50 interpretations. Pairwise Pearson's correlation coefficients between all 50 interpretation maps were computed and then averaged.

765

766 Data availability

The analyses and figures of the manuscript can be replicated with the scripts openly available at https://github.com/EtienneTho/privavox

769

The Spectro-Temporal Modulations (STMs) model adapted from the NSL toolbox (33) is

771 available at: <u>https://github.com/EtienneTho/strf-like-model</u>

772 **References**

773

- 1. Y. S. Bin, N. S. Marshall, N. Glozier, Sleeping at the Limits: The Changing Prevalence of Short and Long Sleep Durations in 10 Countries. *Am J Epidemiol* **177**, 826–833 (2013).
- 2. S. Wang, M. E. Rossheim, R. R. Nandy, Trends in prevalence of short sleep duration and trouble sleeping among US adults, 2005–2018. *Sleep* (2022)
- 779 https:/doi.org/10.1093/sleep/zsac231.
- 3. D. Leger, C. Stepnowsky, The Economic and Societal Burden of Excessive Daytime
 Sleepiness in Patients with Obstructive Sleep Apnea. *Sleep Med Rev* 51, 101275 (2020).
- 4. V. Bayon, D. Leger, D. Gomez-Merino, M.-F. Vecchierini, M. Chennaoui, Sleep debt and obesity. *Ann Med* **46**, 264–272 (2014).
- 5. A. Smiley, D. King, A. Bidulescu, The Association between Sleep Duration and Metabolic
 Syndrome: The NHANES 2013/2014. *Nutrients* **11**, 2582 (2019).
- 6. F. P. Cappuccio, D. Cooper, L. D'Elia, P. Strazzullo, M. A. Miller, Sleep duration predicts
 cardiovascular outcomes: a systematic review and meta-analysis of prospective studies. *Eur Heart J* 32, 1484–1492 (2011).
- 789 **7**. D. Leger, *et al.*, Sleep, substance misuse and addictions: a nationwide observational survey 790 on smoking, alcohol, cannabis and sleep in 12,637 adults. *J Sleep Res* **31**, e13553 (2022).
- 8. B. Faraut, *et al.*, Immune disruptions and night shift work in hospital healthcare
 professionals: The intricate effects of social jet-lag and sleep debt. *Front Immunol* **13**, 939829
 (2022).
- 9. S. Sabia, *et al.*, Association of sleep duration at age 50, 60, and 70 years with risk of
 multimorbidity in the UK: 25-year follow-up of the Whitehall II cohort study. *Plos Med* 19,
 e1004109 (2022).
- 10. G. Kecklund, J. Axelsson, Health consequences of shift work and insufficient sleep. *Bmj*355, i5210 (2016).
- 11. A. D. Larsen, *et al.*, Night work, long work weeks, and risk of accidental injuries. A registerbased study. *Scand J Work Environ Heal* **43**, 578–586 (2017).
- 12. J. Krajewski, *et al.*, Large Sleepy Reading Corpus (LSRC): Applying Read Speech for
 Detecting Sleepiness in (2016), pp. 1–4.
- 13. J. Krajewski, S. Schnieder, D. Sommer, A. Batliner, B. Schuller, Applying multiple
 classifiers and non-linear dynamics features for detecting sleepiness from speech.
 Neurocomputing 84, 65–75 (2012).

- 14. J. Krajewski, A. Batliner, M. Golz, Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach. *Behav Res Methods* **41**, 795–804 (2009).
- 15. H. P. A. VanDongen, K. M. Vitellaro, D. F. Dinges, Individual Differences in Adult Human
 Sleep and Wakefulness: Leitmotif for a Research Agenda. *Sleep* 28, 479–496 (2005).
- 16. P. A. VanDongen, M. D. Baynard, G. Maislin, D. F. Dinges, Systematic Interindividual
- Differences in Neurobehavioral Impairment from Sleep Loss: Evidence of Trait-Like Differential Vulnerability. *Sleep* **27**, 423–433 (2004).
- 17. A. W. MacLean, G. Cynthia, P. Saskin, J. B. Knowles, Psychometric evaluation of the
 Stanford Sleepiness Scale. *J Sleep Res* 1, 35–39 (1992).
- 18. A. Shahid, K. Wilkinson, S. Marcu, C. M. Shapiro, STOP, THAT and One Hundred Other Sleep Scales. 209–210 (2011).
- 19. E. Hoddes, V. Zarcone, H. Smythe, R. Phillips, W. C. Dement, Quantification of Sleepiness: A New Approach. *Psychophysiology* **10**, 431–436 (1973).

20. J. Durmer, D. Dinges, Neurocognitive Consequences of Sleep Deprivation. *Semin Neurol* 25, 117–129 (2005).

- 21. D. L. Arand, M. H. Bonnet, Chapter 26 The multiple sleep latency test. *Handb Clin Neurology* **160**, 393–403 (2019).
- 22. J. Lim, D. F. Dinges, Sleep Deprivation and Vigilant Attention. *Ann Ny Acad Sci* **1129**,
 305–322 (2008).
- 23. C. Bougard, *et al.*, Motorcycling performance and sleepiness during an extended ride on a dynamic simulator: relationship with stress biomarkers. *Physiol Meas* **41**, 104004 (2020).
- 24. T. L. D. H. Editorial, Do I sound sick? *The Lancet Digital Health*, e534 (2021).
- 25. G. Fagherazzi, A. Fischer, M. Ismael, V. Despotovic, Voice for Health: The Use of Vocal Biomarkers from Research to Clinical Practice. *Digital Biomarkers* **5**, 78–88 (2021).
- 26. G. O. Morris, H. L. Williams, A. Lubin, Misperception and Disorientation During Sleep
 Deprivation. *M Archives Gen Psychiatry* 2, 247–254 (1960).
- 27. Y. Harrison, J. A. Horne, Sleep Deprivation Affects Speech. Sleep 20, 871–877 (1997).
- 28. M. Huckvale, A. Beke, M. Ikushima, Prediction of Sleepiness Ratings from Voice by Man
 and Machine. *Interspeech 2020*, 4571–4575 (2020).
- 29. T. L. Nwe, H. Li, M. Dong, Analysis and detection of speech under sleep deprivation.
 Interspeech 2006, paper 1934-Wed2BuP.15-0 (2006).

30. B. W. Schuller, *et al.*, The INTERSPEECH 2019 Computational Paralinguistics Challenge:
Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. *Interspeech 2019*,

839 **2378–2382 (2019)**.

- 31. B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, The INTERSPEECH 2011
 speaker state challenge. *Interspeech 2011*, 3201–3204 (2011).
- 32. B. Günsel, C. Sezgin, J. Krajewski, SLEEPINESS DETECTION FROM SPEECH BY
 PERCEPTUAL FEATURES. 2013 leee Int Conf Acoust Speech Signal Process, 788–792
 (2013).
- 33. V. P. Martin, J.-L. Rouas, P. Thivel, J. Krajewski, Sleepiness detection on read speech
 using simple features. 2019 Int Conf Speech Technology Human-computer Dialogue Sped 00,
 1–7 (2019).
- 34. 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 Digital Heal* 3, 686068
 (2021).
- 35. T. Chi, P. Ru, S. A. Shamma, Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* **118**, 887–20 (2005).
- 36. K. Patil, D. Pressnitzer, S. Shamma, M. Elhilali, Music in our ears: the biological bases of
 musical timbre perception. *PLoS Comput Biol* 8, e1002759 (2012).
- 37. E. Thoret, B. Caramiaux, P. Depalle, S. McAdams, Learning metrics on spectrotemporal
 modulations reveals the perception of musical instrument timbre. *Nat Hum Behav* 5, 369–377
 (2021).
- 38. N. Mesgarani, S. Shamma, SPEECH PROCESSING WITH A CORTICAL
 REPRESENTATION OF AUDIO. 2011 leee Int Conf Acoust Speech Signal Process Icassp 1,
 5872–5875 (2011).
- 39. E. Thoret, T. Andrillon, D. Léger, D. Pressnitzer, Probing machine-learning classifiers using
 noise, bubbles, and reverse correlation. *J Neurosci Meth* 362, 109297 (2021).
- 40. D. Léger, *et al.*, "You look sleepy..." The impact of sleep restriction on skin parameters and
 facial appearance of 24 women. *Sleep Med* 89, 97–103 (2022).
- 41. N. Ding, *et al.*, Temporal modulations in speech and music. *Neuroscience and Biobehavioral Reviews*, 1–7 (2017).
- 42. O. Joly, F. Ramus, D. Pressnitzer, W. Vanduffel, G. A. Orban, Interhemispheric differences
 in auditory processing revealed by fMRI in awake rhesus monkeys. *Cereb. Cortex* 22, 838–
 853 (2012).
- 43. A. del Bimbo, *et al.*, Opensmile. *Proc. 18th ACM Int. Conf. Multimedia*, 1459–1462 (2010).
- 44. G. Varoquaux, Cross-validation failure: Small sample sizes lead to large error bars. *NeuroImage* **180**, 68–77 (2018).
- 45. F. Gosselin, P. G. Schyns, RAP: a new framework for visual categorization. *Trends Cogn Sci* **6**, 70–77 (2002).

- 46. K. L. Knutson, K. Spiegel, P. Penev, E. V. Cauter, The metabolic consequences of sleep deprivation. *Sleep Med Rev* **11**, 163–178 (2007).
- 47. A. J. Krause, *et al.*, The sleep-deprived human brain. *Nat Rev Neurosci* 18, 404–418
 (2017).
- 48. M. W. L. Chee, J. C. Tan, Lapsing when sleep deprived: Neural activation characteristics of resistant and vulnerable individuals. *Neuroimage* **51**, 835–843 (2010).
- 49. B. T. T. Yeo, J. Tandi, M. W. L. Chee, Functional connectivity during rested wakefulness predicts vulnerability to sleep deprivation. *Neuroimage* **111**, 147–158 (2015).
- 50. T. Kaufmann, *et al.*, The brain functional connectome is robustly altered by lack of sleep. *Neuroimage* **127**, 324–332 (2016).
- 51. G. Bernardi, *et al.*, Neural and Behavioral Correlates of Extended Training during Sleep
 Deprivation in Humans: Evidence for Local, Task-Specific Effects. *J Neurosci* 35, 4487–4500
 (2015).
- 52. C.-S. Hung, *et al.*, Local Experience-Dependent Changes in the Wake EEG after
 Prolonged Wakefulness. *Sleep* 36, 59–72 (2013).
- 53. B. Faraut, T. Andrillon, M.-F. Vecchierini, D. Leger, Napping: A public health issue. From
 epidemiological to laboratory studies. *Sleep Med Rev* 35, 85–100 (2017).
- 54. B. Faraut, V. Bayon, D. Léger, Neuroendocrine, immune and oxidative stress in shift
 workers. *Sleep Med Rev* 17, 433–444 (2013).
- 55. F. Sauvet, *et al.*, Effect of acute sleep deprivation on vascular function in healthy subjects. *J Appl Physiol* **108**, 68–75 (2010).
- 56. E. A. Dolsen, A. D. Crosswell, A. A. Prather, Links Between Stress, Sleep, and Inflammation: Are there Sex Differences? *Curr. Psychiatry Rep.* **21**, 8 (2019).
- 57. V. P. Martin, B. Arnaud, J.-L. Rouas, P. Philip, Does sleepiness influence reading pauses in hypersomniac patients? *Speech Prosody* 2022, 62–66 (2022).
- 58. K. R. Baykaner, M. Huckvale, I. Whiteley, S. Andreeva, O. Ryumin, Predicting Fatigue and
 Psychophysiological Test Performance from Speech for Safety-Critical Environments. *Front. Bioeng. Biotechnol.* 3, 124 (2015).
- 59. M. Ghassemi, L. Oakden-Rayner, A. L. Beam, The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digital Heal* **3**, e745–e750 (2021).
- 60. E. Thoret, P. Depalle, S. Mcadams, Perceptually Salient Regions of the Modulation Power
 Spectrum for Musical Instrument Identification. *Front. Psychol.* 8, 1–10 (2017).
- 907 61. B. N. Pasley, *et al.*, Reconstructing Speech from Human Auditory Cortex. *Plos Biol* **10**,
 908 e1001251 (2012).

909 62. R. F. Murray, Classification images: A review. *J Vision* **11**, 2–2 (2011).

910

911

912 Acknowledgments

- 913 Author ET was supported by grants ANR-16-CONV-0002 (ILCB), ANR-11-LABX-0036
- 914 (BLRI) and the Excellence Initiative of Aix-Marseille University (A*MIDEX) (ET). Author
- 915 DP was supported by grants ANR-19-CE28-0019-01 and ANR-17-EURE-0017. Author
- 916 TA was supported by a Human Frontier Science Program Long-Term Fellowship
- 917 **(T000362/2018-L)**.
- 918

919 **Author contributions:**

- 920 Conceptualization: ET, TA, DL, DP
- 921 Methodology: ET, TA, CG, DL, DP
- 922 Investigation: ET, TA, CG, DL, DP
- 923 Visualization: ET, TA, DL, DP
- 924 Supervision: DP, DL
- 925 Writing—original draft: ET, TA, DL, DP
- 926 Writing—review & editing: ET, TA, DL, DP
- 927

928

Competing interests: "All other authors declare they have no competing interests."

929

Data and materials availability: The analyses and figures of the manuscript can be replicated
 with the scripts openly available at https://github.com/EtienneTho/privavox The Spectro Temporal Modulations (STMs) model adapted from the NSL toolbox (33) is available at:
 https://github.com/EtienneTho/privavox