Sleep deprivation measured by voice analysis

Etienne Thoret[1,4,5], Thomas Andrillon[2,3], Caroline Gauriau[2], Damien Léger[2, *], Daniel Pressnitzer[1, *]

[1] Laboratoire des systèmes perceptifs, Département d'études cognitives, École normale supérieure, PSL University, CNRS, 75005 Paris, France.

[2] APHP, Hôtel Dieu, Centre du Sommeil et de la Vigilance, 75004 Paris, France & Université Paris Cité, VIFASOM, ERC 7330 Vigilance Fatigue Sommeil et Santé Publique, 75006 PARIS, France.

[3] Turner Institute for Brain & Mental Health and School of Psychological Sciences, Monash University, Melbourne 3168, Australia.

[4] Aix-Marseille University, CNRS, UMR7061 PRISM, UMR7020 LIS, Marseille, 13009, France

[5] Institute of Language Communication and the Brain, Marseille, France

[*] Joint last authors.

Corresponding author: *etiennethoret@gmail.com*

Keywords: Sleep – Voice – Biomarkers - Auditory modeling - Spectrotemporal modulations – Machine learning – Explainable AI (XAI)

Abstract

Using our voice represent an exquisitely intricate act, recruiting a host of cognitive and motor functions. As such, the voice is bound to reflect many aspects of the internal state of the speaker: personality, infections, stress, emotions. Here, we investigate whether sleep deprivation in otherwise normal and healthy persons can be detected through machine-learning analysis of vocal recordings. In contrast to previous approaches, we use fully generic acoustic features, derived from auditory-inspired models of sound processing, together with recently-developed machine learning interpretation techniques. Our results show that sleep deprivation can be accurately detected from generic acoustic features of vocal recordings. Two main different types of features were impacted by sleep deprivation: one related to speech rhythms, the other related to the timbre of the voice. We speculate that these features reflect two distinct physiological processes: the cognitive control of speech production and an inflammatory effect of the vocal apparatus. Crucially, the relative balance of these two effects varied widely for each individual, suggesting that the voice may be used as a "sleep stethoscope" to better understand the variety of idiosyncratic responses to sleep deprivation. Moreover, the method we outline is fully general and could be applied to the future investigation of any type of vocal biomarkers using machine-learning techniques.

Introduction

In the last decade or so, insufficient sleep in adults with sleep debt has become an increasing prominent public health issue, with one third of the adult population sleeping less than six hours per night (1–3). This chronic sleep debt in adults is associated with an increased risk of chronic disease such as obesity, type 2 diabetes, cardiovascular diseases, inflammation, addictions, accidents and cancer (4–8). Short sleep duration and sleep debt also increase the risk of developing multiple comorbidities (9). Moreover, more than one worker out of five operates at night and suffers from a high level of sleep deprivation (10). Such a level of sleep deprivation can cause accidents in the workplace or when driving (11) with a severe impact on quality of life, sociability and economics.

Currently, there are several techniques aiming to measure sleep deprivation and its associated physiological consequences. From the outset, a core issue needs to be considered: the most appropriate measure of sleep deprivation and its associated adverse consequences is far from being obvious. Sleep deprivation may be simply assessed in terms of the loss of sleep time, but, remarkably, the impact of a given amount of sleep deprivation varies massively across individuals. In laboratory settings where the amount of deprivation could be precisely controlled, it has been shown that up to 90% of the variance in cognitive performance was related to individual traits and not to the actual time spent asleep (12, 13). Sleep deprivation may also be measured with subjective sleepiness, which participants can explicitly report using sleepiness scales (14–16). However, subjective sleepiness is influenced by other factors than sleep deprivation, such as the time of the day, motivation, or stress. Besides, it is not at all obvious whether reported subjective sleepiness captures the full physiological impact of sleep deprivation, given the variety of processes involved (17). Given the practical importance of quantifying precisely how sleep deprivation affects behavior (e.g., for truck or train drivers), objective methods have also been developed. The multiple sleep latency test (18), the gold standard in clinical settings, uses electro-encephalography (EEG) to estimate sleep latency (*e.g.* the amount of time to go from wake to sleep) along five successive naps sampled every two hours during daytime. The psychomotor vigilance test (19), often used in research settings, tests for the ability to respond quickly to infrequent stimuli, with slower reaction times being used as a marker of attentional lapses. More recently, new approaches have attempted to leverage the concentration ok key molecules in the urine, saliva or breath to quantify sleepeiness (20). Although these objective methods are complementary to subjective

reports, they are often costly, time consuming, or difficult to deploy outside of laboratories. So, whereas there exists cheap and fast objective measures for other causes of temporary cognitive impairment, such as alcohol or drug abuse, there is currently no established means to estimate sleep deprivation effects at the individual level in real-life settings.

If sleep deprivation could be detected reliably through voice recordings, this would provide a quick, non-invasive, and cost-effective objective measure of sleep deprivation. Indeed, because the voice is easy to record with off-the-shelf equipment, there is a growing interest in finding vocal biomarkers to diagnose a variety of medical conditions (The Lancet Digital 21, 22). An early report suggesting an impact of sleep deprivation on the voice was provided by Morris et al. (23). Based on subjective evaluations by the experimenters of free speech produced by sleep deprived participants, a slowing down of speech and a "flatness of the voice" were noted after deprivation. These observations were extended by Harrison and Horne (24), who used a word fluency task and asked raters, blind to the amount of deprivation, to evaluate speech quality on five scales (intonation, errors, volume, fatigue, pace average). Effects of deprivation were found on the "intonation" and "fatigue" scales. More recently, an experiment using more raters on a large recorded database found that indeed, raters could detect sleepy versus non sleepy voices with an accuracy above 90% (25), even though the precise degree of self-reported sleepiness was not well estimated by raters. So, it does seem that there are acoustic cues in the voice that reflect sleep deprivation and/or sleepiness.

In recent years, the field of machine learning has taken up the challenge to automate the detection of sleep deprivation or sleepiness from the voice. In an early study (26), sleep deprivation was inferred with high accuracy from vocal recordings (86%) but it should be noted that the deprivation was extreme, consisting of 60 hours without sleep in a setting involving military personnel, with debatle applicability for mild sleep deprivation. Two "computational paralinguistic challenges" have since been launched, with sub-challenges aimed at assesing sleepiness from vocal recordings (27, 28). We will not review all of the entries to these challenges, as they are quite technical in nature. To summarize, all of them used the same framework: i) selection of a set of acoustic features, such as pitch, spectral and cepstral coefficients, duration estimates, and functionals of those features; ii) dimensionality reduction of the feature set; iii) supervised learning of target classification using various machine learning techniques, such as support vector machines or neural networks. The best results varied depending on the challenge. Subjective

sleepiness proved difficult to predict (25), but the binary categorization of sleepy versus non-sleepy voices could be achieved with high accuracy (over 80%) in the best performing classifiers (29).

The computational framework described above will be familiar and effective for many machine learning problems, but it has two major limitations in a neuroscientific perspective. First, the initial selection of features is based on a somewhat arbitrary choice. Often, the choice of features was guided by the "voice flatness" hypothesis (23, 24). However, there is no guarantee that other, perhaps more subtle acoustic markers of sleep deprivation or sleepiness may not have been overlooked. Second, the acoustic features discovered by the classifier to achieve its task are not necessarily interpretable and difficult to relate to plausible physiological mechanisms that could mediate the effects of sleep deprivation on the voice (30). Interestingly, the best-performing system used a carefully hand-crafted small feature set inspired from auditory processing models, suggesting that "perceptual" features may be a promising route for sleepiness detection in the voice (29). A more recent study has again attempted to focus on "simple" acoustic descriptors for one of the databases of the paralinguistic challenge, with the explicit aim to facilitate interpretation (31). 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 important ways. First, we use our own vocal database, which has been collected in a controlled laboratory setting where the amount of sleep deprivation could be precisely controlled. Vocal recordings were obtained from reading out loud the same texts for all participants, in random order across participants, thus avoiding biases confounding *e.g.* participant identity and sleep deprivation observed in previous databases (25, 32). Second, we use a fully generic acoustic feature set, derived from an established model of auditory processing (33). Our audio input representation is based on so-called *spectro-temporal modulations* (STMs). To compute STMs, a sound is first split into separate frequency bands (in our case on a log-frequency scale) to produce a cochleagram, simulating peripheral auditory processing. The joint modulations over time and frequency in the cochleagram are then estimated with an analysis akin to a 2D-Fourier transform. While the STM representation was initially motivated by neurophysiological results, it has been found to be sufficiently rich to successfully support various machine-learning tasks like musical instruments classification (34), timbre perception (34, 35), or speech detection and enhancement (36). Third, we apply a recently-

developed machine-learning technique to interpret the cues discovered by our classifier (37). This technique, similar in spirit to the reverse correlation method broadly used in neuroscience and psychophysics, identifies the parts of the input representation that have the most weight in the classifiers' decisions. Finally, by fitting classifiers to each participants, we can hope to uncover the individual physiological factors underlying the large and as yet unexplained variability observed in the responses to mild sleep deprivation in normal healthy adults.

As will be seen, our results show that sleep deprivation can be detected in generic acoustic features of vocal recordings, with two main different acoustic cues that vary across individuals: one related to speech rhythms, the other related to the timbre of the voice. Speech rhythm mainly relies on the "melody of speech", i.e., the slow temporal variations, and timbre of the voice mainly relies on the spectral content of the voice signal. A moderate correlation with subjective sleepiness reports is observed for rhythmic cues only. On the basis of these acoustic signatures, we propose that the classifiers distinguish two distinct neurophysiological effects of sleep deprivation: a cognitive control effect on speech rhythms, which is partially reflected in subjective sleepiness reports, and an inflammatory effect of the vocal apparatus, which will impact timbre but may escape subjective report techniques. Crucially, the relative balance of these two effects varies widely for each individual, suggesting that the voice may be used as a "sleep stethoscope" to better understand the variety of responses that each and everyone of us displays after sleep deprivation.

Results

Twenty-two healthy women between 30-50 years of age (42.7 ± 6.8) were sleep deprived during a controlled laboratory protocol. An all-female experimental group was chosen because the current experiment took place in parallel with a dermatology study (38), and also because such a choice was expected to homogenize the vocal pitch range across participants. After a first "Control night" spent in the laboratory, participants were restricted to no more than 3 hours of sleep per night during two subsequent "Restriction nights", also monitored in the laboratory. Sleep restriction is a more ecological manipulation of sleep time than total sleep deprivation. Vocal recordings were obtained throughout the protocol, during reading sessions sampled at different times of the day. These reading sessions occurred either: i) right after the control night (no sleep deprivation); or ii) right after the second restriction night (see Methods for details). All participants read the same excerpts from a French classic novel ("Le Comte de Monte-Cristo"). The order of the excerpts was randomized across sessions for each participant to avoid a confound with deprivation. In total, our database consists of 22 healthy participants producing about half an hour of vocal recordings (M=31min, SD=5min) evenly split between before and after two nights of mild sleep deprivation.

Subjective sleepiness reports

[Insert Figure 1 about here]

Sleepiness was also self-reported by participants by means of a Stanford Sleepiness Scale questionnaire (16) sampled at different times during the day (see Methods). Figure 1A shows the distributions of SSS ratings. On average, sleep deprivation had an effect on self-reported sleepiness: sleepiness was low right after the control night, but increased after the deprivation nights. This was confirmed by an ANOVA on the SSS, with factors Day (2 levels, before and 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 will now consider average SSS values for all Times of reports per Day, to focus on the effect of sleep deprivation. Figure 1B illustrates the data aggregated in that way, with individual changes in sleepiness now identified across the control and sleep deprived day. A remarkable individual variability was

obvious in reported sleepiness. Note that this was in spite of our precise control of the amount of sleep deprivation, which was equated across all participants. Even so, some 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, low or high, as measured before deprivation. This striking variability is in fact consistent with previous observations involving objective measures of sleep deprivation (13). It also confirms that vocal biomarkers of sleep deprivation should be investigated at the individual level.

Spectro-Temporal Modulations of speech before and after sleep deprivation

We computed STMs for all speech recordings. At each moment in time, STMs were computed along the dimensions of *frequency*, *rate*, and *scale*. The *frequency* dimension, in Hz, reflects the spectral content of the sound. The *rate* dimension, in Hz, reflects the modulations in sound amplitude in the time domain. Slow modulations have low rates, whereas fast modulations have high rates. Positive rates indicate temporal modulations coupled with downward changes in frequency, whereas negative rates indicate temporal modulations coupled with upward changes in frequency. The *scale* dimension, in cycle per octave, reflects modulations in the spectral domain. Sounds with fine spectral envelopes have high scale values, while sounds with relatively flat spectral shapes have low scale values. For speech, the dominant rates are between 2 Hz and 8 Hz (39), while dominant scales, presumably to the harmonic structure of vowels, are around 2 cyc/oct (40).

[Insert Figure 2 about here]

The full STMs have four dimensions of time, frequency, rate, and scale. To have a first look at the gross effect of sleep deprivation on acoustic features, we averaged the STMs along the time dimension, for all recordings obtained before (Fig. 2A) and after deprivation (Fig. 2B). Average STMs before and after deprivation were qualitatively similar. The rate-scale projections showed that, unsurprisingly, high energy in the STMs was focused in regions associated to speech (36). The frequency-rate projection simply showed the average spectrum of our vocal recordings.

To further highlight the candidate acoustic differences caused by deprivation, we subtracted STMs before and after deprivation (Fig. 2C for the population-level results, Fig. S1 for individual-level results). At the population level, maximal differences in the rate-scale projection were less than 3%, while differences up to 11% were observed in the frequency-rate projection.

At the subject level, differences in the rate-scale projection were around 24.68% on average (SD=6), while differences up to 40.83% on average (SD=12) were observed in the frequency-rate projection. Larger differences seems therefore observable at individual level but there is no obvious structure to the differences: they appear noisy and do not necessarily match the STM regions of high energy in speech (see Supplemental Figure S1). Crucially, it is unclear at this point whether such raw acoustic differences are meaningful, for instance compared to the within- and across-participant variability. To address this point in a principled manner, and with the aim to predict sleep deprivation from the voice, we now turn to machine-learning.

Machine-learning classification of speech before and after sleep deprivation

A standard machine-learning pipeline was used to identify sleep deprivation in vocal recordings (30, 34, 41). For all recordings, STMs were computed and used as the input feature space. The dataset was first transformed into random train/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 for more details). 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 because of the relatively modest size of our dataset, but also because SVMs have been shown to outperform more complex classifiers in similar tasks (30). The performance of the SVM was evaluated with Balanced Accuracy (BAcc, see Methods).

[Insert Figure 3 about here]

The classifiers' balanced accuracies are shown in Figure 3. At the population level, the classifier was able to detect sleep deprivation well above chance (BAcc, M=.77, SD=.01). This is on par with the state of the art (29, 31). Interestingly, and as expected from the sizeable individual variability observed in the SSS reports, the same machine-learning pipeline was more accurate when applied at the individual level (BAcc, M=.85, SD=.09). In this case, the classification for two participants displayed relatively poor BAccs, whereas about half of the remaining participants displayed BAccs above .9, outperforming the state of the art and matching human performance on a similar task (25). This shows, for the first time, 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, at the population level and at the individual level.

Vocal biomarkers of sleep deprivation

To gain insight about the nature of the acoustic features distinguishing speech before and after sleep deprivation, we probed the trained classifiers using a recent interpretation technique based on reverse correlation (37). Briefly, the technique consists in randomly perturbing the input to the trained classifier over thousands of trials and then averaging all of the noisy representations leading to correct classification in order to identify the portion of the input that participates the most to the classifier's performance. The input representation was perturbed using additive noise in the PCA-reduced feature space (42). Averaging all masks leading to a correct classification decision revealed, in our case, the discriminative features of a voice after deprivation compared to before deprivation (for details, see Methods and 37).

As a preliminary step, we evaluated the consistency of the interpretation masks. Because of our cross-validation technique, 50 classifiers were fitted either for the whole dataset for the population-level classifier or for each participant's classifier. To check internal consistency, we computed the pairwise Pearson's correlation coefficients between all 50 interpretation maps. At the populationl-level, this "consistency correlation" was low (r(22527): M = .20, SD=.34, $BF_{10}>2^{16}$) which is coherent with the large variability suspected across listeners. At the participant-level, however, consistency correlations were very high (r(22527): M=.91, SD=.06, min=.73; All BF_{10} are high > 2^{16}). Furthermore, because individual classifiers varied in accuracy, we could check whether the consistency of the interpretation improved with accuracy. As expected, the correlation between BAccs and consistency correlation was strong (r(20)=.71, p=.0003, BF_{10}=130.67). These consistency results suggest that caution should be applied when considering population-level interpretations, but that individual results are robust and can be interpreted.

Figure 4 shows the interpretation maps for the population-level classifier. Maps should be read as follows: *red* areas correspond to STM features where the *presence* of energy is associated with sleep deprivation for the classifier, whereas *blue* areas represent STM features where the *absence* of energy is associated to sleep deprivation for the classifier. For the population-level map, the rate-scale projection resembles the raw difference before and after deprivation, although less noisy, whereas the frequency-rate projection does not match such raw acoustic differences (compare with Fig. 2C). As these population-level interpretation are not robust, we simply show them for illustrative purposes and refrain from further description of their features.

Figure 5A shows all individual classifiers on the rate-scale projection, ordered along increasing accuracy (BAcc) of the corresponding classifier. We chose to interpret in priority the rate-scale projections, as is done in most speech applications (36), but the frequency-rate projections are provided as Supplemental Figure S3. The main feature of the results is the striking diversity of the individual maps, which is not related to classifier accuracy in any obvious manner. For some participants, sleep deprivation was detected through a reduction in energy over a range of STM features (blue color), consistent with a "flattening" of speech modulations. But the opposite was also observed for other participants (red color). Moreover, the details of the discriminative features also varied across participants. As shown before, these details are robust and warrant interpretation.

To get a better understanding of this variability across individual maps, we performed a PCA on the maps themselves, which we will term interpretation-PCA for clarity. A first interpretation-PCA dimension explained 35.9% of the variance, while a second dimension explained 24.2% of the variance. There is a drop for all other dimensions (N=3) which explain less that 13% of the variance, see Supplemental Figure S4. Participants ordered on the first two interpretation-PCA dimensions are shown in Figure 5B. We computed the variance of all STM features along each interpretation-PCA dimension, to visualize the features that distinguished the interpretation maps along these main axes of variation. Results are shown in Figure 5C. The features defining the first interpretation-PCA dimension were clustered between rates of about 2 Hz to 8 Hz, which is exactly the amplitude modulation range corresponding to speech prosody and syllabic rate (39). This shows that the amplitude modulation characteristics of speech was affected by sleep deprivation. Importantly, depending on the individual, the classifiers used the presence *or* absence of energy around these rates to detect sleep deprivation. This shows that while some participants spoke in a "flattened" voice after deprivation, consistent with classic hypotheses (23, 31) others instead spoke in a more "animated" voice after deprivation. The features defining the second interpretation-PCA dimension clustered at very low rates and covered a broad scale range, peaking at about 2 cyc/oct. This corresponds to long-term spectral characteristics of speech and vowel sounds. In speech, such timbre-like features are determined 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 cords, resonators impose formants that impact the timbre of vowels and other speech sounds.

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 the individual 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. 6A. There was no significant correlation between BAccs and SSS reports ($r(20)=.32$, $p=.14$, $BF_{10}=.47$), 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 SSS reports variability. In particular, we reasoned that the prosodic and rhythmic changes 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-monitoring their own speech rate and intonation. In contrast, the timbre cues captured by the second interpretation-PCA dimension could be more subtle and remain implicit. Results are shown in Fig. 6B & 6C. Consistent with our hypothesis, we observed a moderate but significant correlation between the location of participants on the first interpretation dimension and sleepiness reports ($r(20)=-.44$, $p=.03$, $BF_{10}=1.34$). In contrast, the location of participants on the second interpretation dimension did not show any significant correlation with sleepiness reports ($r(20)=.19$, $p=.38$, $BF_{10}=.23$).

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 (R^2: $M=.29$, $SD=.18$) 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.

Discussion

Summary of findings

We ran a sleep deprivation protocol with normal and healthy participants, collecting subjective reports of sleepiness plus vocal recordings before and after deprivation. After two nights of mild sleep deprivation, subjective sleepiness increased on average, although with striking individual differences—including some participants reporting decreases in subjective sleepiness after deprivation. Nevertheless, sleep deprivation could be detected accurately in all participants by means of machine-learning analysis of their vocal recordings. Classification was most accurate at the individual level, with 85% balanced accuracy on average. Importantly, such a classification was based on a fully generic auditory representation of sound. This allowed us to interpret the discriminative features discovered by classifiers to detect sleep deprivation. Two broad 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).

As discussed below, we propose that these two classes of features titrate two independent physiological processes that are known to be triggered by sleep deprivation and that could plausibly impact vocal production: impaired cognitive control of the speech production process revealed by temporal modulations, and an inflammatory response of the vocal tract revealed by spectral modulation. If, as our results suggest, the relative balance between the two processes can assessed through the voice, this opens up clinical perspectives for a future rapid and cheap assessment of the individual neurophysiological consequences of sleep deprivation, akin to a "sleep stethoscope". Moreover, because of its generality, the method we developed could be easily extended to interpret any vocal biomarker in a broad range of conditions.

Physiological interpretation of the vocal biomarkers of sleep deprivation

Our individual classifiers 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 relate these 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 has been robustly found as characteristic of speech across a variety of languages, to the extent that they have been described as "universal rhythmic properties of human speech" (39). According to the same authors, such a universal rhythm is imposed by the biomechanical constraints of the vocal apparatus and by the neurodynamics of its control and perception systems. We suggest that the changes in speech rhythms observed after sleep deprivation result from a temporary impairment of the cognitive control of the speech production process. Sleep deprivation impacts cognitive function (17), presumably through changes in glucose consumption in frontal and motor brain regions (43, 44). Accordingly, previous studies showed lower activity in the dorsolateral prefrontal cortex and in the intraparietal sulcus in cognitive tasks requiring attention, with large inter-individual variability (45). 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 (44, 46, 47). Finally, extended wakefulness has been associated with an increase in the intrusion of sleep-like patterns of brain activity in wakefulness(48, 49). 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 (23, 24), whereas others unexpectedly responded with speech over-modulation and instead produced an "animated voice".

The spectral modulation changes detected by our classifiers were consistent with changes in the timbre of speech sounds, and in particular vowel sounds (33, 36, 40). Such sounds acquire their distinctive spectral envelopes by means of the resonances of the vocal tract, including the throat and nasal cavities. We suggest that inflammation of the throat and nose may be 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 sleep deprivation (8, 50, 51), so is plausible in our protocol that included two nights of mild sleep deprivation. In terms of mechanisms, sleep restriction and deprivation disturb the normal secretion of hormones like cortisol or testosterone, and is associated with increased rates of interleukin-6 and CRP as assessed on salivary samples in normal subjects. This inflammatory response could be linked to an elevated

blood pressure following sleep deprivation (52) and could affect the vocal tract and plausibly impact the spectral envelope of speech.

Limitations of the present study

There are both technical and conceptual limitations to the present study. We chose to use a controlled protocol to precisely equate sleep deprivation in our participants, but this costly protocol came at the expense of a relatively small dataset compared to the online databases used by machine-learning challenges (27, 28). This removed biases in the database, such as associating the identity of speakers with the amount of sleep deprivation (25), but it also limited our choice of possible machine-learning techniques to perform the classification. We thus used an SVM classifier, and not potentially more powerful deep-learning architectures. We note however that in the studies that compared SVMs with other classifiers types, SVM performed best, including in state-of-the-art studies (29–31). In any case, the interpretation method we used could be applied to any kind of classifier (37), including more complex ones. We believe that the robustness of the interpretation maps, shown by our various consistency checks, makes it unlikely that our qualitative conclusions depend on the details of the classifiers.

The feature set we used was a generic and familiar auditory representation, which is a major differences with previous machine-learning oriented studies. On the one hand, some studies were fully data-driven and extracted features from thousands of speech descriptors. The resulting features were often difficult to interpret. On the other hand, there were also studies using a small set of features, but these features were carefully hand-crafted and 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 were from a generic representation, they remained interpretable. A clear limitation is that we did 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 (29, 30, 53). Given the high accuracy of our classifiers, we believe that such omission was compensated by the richness of the STM representation. Pitch and pitch functionals will in fact be indirectly reflected in the STM, which analyses sounds over a range of temporal scales simultaneously.

Another methodological consideration is that we do not have any independent measure of either the cognitive factor nor of the inflammatory factor that, we speculate, are the

neurophysiological bases of our vocal biomarkers. The cognitive factor could be assessed with objective behavioral measures, such as the psychomotor vigilance test (19), or with brain imaging data (43, 44). The inflammatory factor could be assessed by biological analyses of *e.g.* cortisol in the saliva (8, 51). Because we have not gathered such measurements, we can only argue that both minor cognitive dysfunction and inflammation effects are extremely likely for our participants as a group. The present study being the first one to suggest that such factors may be measured at the individual from voice biomarkers, it is left for future investigations to correlate them with other objective markers.

On a conceptual level, we finally need to raise a basic but inescapable limitation of any study of sleep deprivation. Sleep deprivation may be objectively quantified, as we did, by the amount of sleep available to each individual. However, as has been repeatedly pointed out and again observed 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 effects of sleep deprivation. Subjective reports may capture explicit feelings of fatigue but be blind to more elusive effects such as inflammation. With objective measures, which are by necessity indirect, there is an issue with interpreting negative outcomes. In our case for instance, how to interpret a relatively poor accuracy for a sleep deprivation classifier, such as we observed for two participants? It cannot be decided whether this poor accuracy shows that sleep deprivation had no effect on these participants, or that sleep deprivation had effects that were not expressed in the voice, or that the classifiers failed simply for technical reasons. Measuring multiple markers of sleep deprivation, including the novel ones we suggest, and incorporate them into a holistic model of the neurophysiological effects of sleep deprivation seems to be a promising way forward.

Perspectives

Keeping these limitations in mind, the demonstration of vocal biomarkers for sleep deprivation could have major clinical implications. Subjective sleepiness reports may not capture the whole effect of a lack of sleep. Moreover, such reports rely on the honest cooperation of participants, which is not a given if self-reports of excessive sleepiness can have negative work-related or financial consequences for the individual. Objective correlates of sleepiness exist to alleviate these concerns (18, 19), but vocal biomarkers would represent a considerably cheaper and faster alternative, requiring no specialized equipment and increasing their practicality for real-

life clinical assessment. Crucially, our technique also goes beyond the simple binary detection of sleep deprivation: thanks to the application of interpretability techniques (37), we suggest that different neurophysiological processes related to sleep deprivation may be untangled through the voice alone.

Another potential appeal of teasing apart different effects of sleep deprivation at the individual level is to try to understand the well-documented variability in responses to the same amount of deprivation. Vocal biomarkers could be used to titrate the amount of cognitive vs inflammatory responses to sleep deprivation, to objectivate the individual trait assumed to underlie responses to deprivation (12). Such measures could in turn be used to design interventions tailored to each individual and situation, if the effects of sleep deprivation needed to be temporarily alleviated for 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 to understand the reason for a diagnostic (54). For application to real-life settings, it is particularly interesting to have identified features that correlate with subjective sleepiness but also features that do not correlation with such subjective assessments as one of the biggest danger of sleep loss is the partial agnosia for one's own sleepiness.

Finally, it is useful to point out that the methodological pipeline we introduced is fully generic, as the audio features representation we used is itself generic and the interpretation method we developed be applied to any classifier, irrespective of its architecture or complexity. 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 (21, 22).

Experimental methods

Experimental design.

A group of twenty-four healthy women between 30-50 years old (42.7 ± 6.8) took part in the experiment. This study was part of a demartological 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 ≤ 10), and were not under any medical treatment (exclusion criteria).

Before the experiment, participants wore an actigraph for 7 days and were instructed to maintain a regular sleep-wake behavior with their usual 7-8 h of sleep (i.e., in bed from 23:00-01:00 until 07:00-09:00). The compliance with these recommendations was verified through the actigraphic recordings (MW8, CamTech; UK) that were inspected by the research team at the participant's arrival the morning before the first night of sleep restriction (day 1). No sleep episodes were detected outside of the scheduled experimental time in bed (see 38 for details). The protocol lasted for 3 days (day 1: before sleep restriction; day 2: during sleep restriction; day 3: after sleep restriction), which included 2 night of sleep deprivation (at the end of day 1 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 usual routine outside the laboratory. After the second sleep-restricted night (day 3), the 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 day 1 (after habitual sleep and before sleep restriction: baseline condition) and day 3 of each 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 at the end of day 1 until their return to the laboratory at the beginning of day 3 at 09:00, two investigators exchanged text messages with the participants at random times during the entire period outside of the laboratory. Text messages were sent throughout the night (except during the period where participants were instructed to sleep, that is between 3 and 6 a.m.). Participants had to respond right after receiving these messages. In case of an absence of response, participants were immediately called on their personal phone. For lunch in the laboratory (day 1 and 3), participants received controlled meals consisting of a

maximum of 2,500 calories/day with a balanced proportion of nutrients (protein, fat, and carbohydrates).

Voice recording

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.

Spectro-Temporal Modulations (STM)

The sound files, initially sampled at 44.1 kHz, were down-sampled to 16 kHz. Spectro-Temporal Modulations (STMs) were computed with our own toolkit (https://github.com/EtienneTho/strf-like-model) which is directly adapted from the standard NSL Toolbox (33). Sounds were processed through a bank of 128 constant-Q asymmetric bandpass filters equally spaced on a logarithmic frequency scale spanning 5.3 octaves, which resulted in an auditory spectrogram, a two-dimensional time-frequency array. The STM were then computed by applying a spectro-temporal modulation filterbank to the auditory spectrogram. We generally followed the procedure detailed in (34), with minor adaptations. A 2D Fourier transform was first applied to the spectrogram resulting in a two-dimensional array, also called Modulation Power Spectrum (MPS) (55) whose dimensions were spectral modulation (scale) and temporal modulation (rate). Then, the STM representation was 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 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, r = ±[.25, .5, 1, 2, 4, 5.70, 8, 11.3, 16, 22.6, 32] Hz. Such a range covers the relevant spectral and temporal modulations of speech sounds as already used in different studies (56). The resulting representation thus corresponds to a 4D matrix with dimensions of time, frequency, scale, and rate.

Classification framework

A standard machine-learning pipeline was used (30, 34, 41). First, at subject/population level the whole dataset was randomly separated into a training set (75%) and a testing set (25%). 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.

We then reduced this high dimensionality of the feature space by means of a principal component analysis (PCA). At the population level, we trained a PCA on the whole dataset and retained the 250 main dimensions, explaining 99% of the variance. We further checked that the exact choice of PCA dimensions did not affect our conclusions, about the performance but also about the interpretation of the classifiers (see Supplemental Figure S2). At the participant level, for each participant we trained a PCA on the data from all other participants, to reduce a possible contamination of the reduced space by peculiarities of the considered participant. We next retained the 30 main dimensions of the PCA. The number of PCA dimensions in this case was chosen empirically, so that the reduced feature space still 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 avoid overfitting. We checked that the exact choice of PCA dimensions did not affect our conclusions, in particular on the interpreted features that are consistent for PCA dimensions above 30.

The PCA spaces were then fed to a support vector machine classifier (SVM) with a gaussian kernel (radial basis function). The training set was used to fit the SVM through an hyperparameter grid-search, using a stratified 5-folds cross-validation. The fitted SVM was then 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). The fitting procedure was repeated 50 times with 50 different train/test sets, at the population and at the individual level. All the classification pipelines from PCA to RBF + SVM are implemented with the sci-kit learn library (Pedregosa et al., 2011).

Interpretation of the classifiers

Each classifier fitted in the study is probed with the reverse correlation technique which provides an interpretation of which features are relevant in the sense of the classifier. 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 to excerpts classified as before sleep restriction and 1 to excerpts recorded after sleep restriction. Here, as the noise were of null average in the feature space, we simply substracted the average values of the noises that led to a classification in the class 'after restriction' with the average values that led to classification in the class 'before restriction' (57). We refer to our method paper for a full description of the method (37). Here, we used the version of the method which consists of pseudo-random noise as perturbation at the input of the system. Pseudo-random noises allow to accurately fool the classifier while using a white noise may implicate complication as the classifier can tend to classify all the stimuli + noise excerpt in only one class. One specificity of this method is that it requires a large number of trials to provide an accurate 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 trials. Each interpretation provides a "interpretation mask" which are composed of positive and 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.

Data availability

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

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

References

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) 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).

7. D. Leger, *et al.*, Sleep, substance misuse and addictions: a nationwide observational survey 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 register-based study. *Scand J Work Environ Heal* 43, 578–586 (2017).

12. 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).

13. 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).

14. A. W. MacLean, G. Cynthia, P. Saskin, J. B. Knowles, Psychometric evaluation of the Stanford Sleepiness Scale. *J Sleep Res* 1, 35–39 (1992).

15. A. Shahid, K. Wilkinson, S. Marcu, C. M. Shapiro, STOP, THAT and One Hundred Other Sleep Scales. 209–210 (2011).

16. E. Hoddes, V. Zarcone, H. Smythe, R. Phillips, W. C. Dement, Quantification of Sleepiness: A New Approach. *Psychophysiology* 10, 431–436 (1973).

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

18. D. L. Arand, M. H. Bonnet, Chapter 26 The multiple sleep latency test. *Handb Clin Neurology* 160, 393–403 (2019).

19. J. Lim, D. F. Dinges, Sleep Deprivation and Vigilant Attention. *Ann Ny Acad Sci* 1129, 305–322 (2008).

20. 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).

21. T. L. D. Health, Do I sound sick? *Lancet Digital Heal* 3, e534 (2021).

22. 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).

23. G. O. Morris, H. L. Williams, A. Lubin, Misperception and Disorientation During Sleep Deprivation. *M Archives Gen Psychiatry* 2, 247–254 (1960).

24. Y. Harrison, J. A. Horne, Sleep Deprivation Affects Speech. *Sleep* 20, 871–877 (1997).

25. M. Huckvale, A. Beke, M. Ikushima, Prediction of Sleepiness Ratings from Voice by Man and Machine. *Interspeech 2020*, 4571–4575 (2020).

26. T. L. Nwe, H. Li, M. Dong, Analysis and detection of speech under sleep deprivation. *Interspeech 2006*, paper 1934-Wed2BuP.15-0 (2006).

27. B. W. Schuller, *et al.*, The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. *Interspeech 2019*, 2378–2382 (2019).

28. B. Schuller, S. Steidl, A. Batliner, F. Schiel, J. Krajewski, The INTERSPEECH 2011 speaker state challenge. *Interspeech 2011*, 3201–3204 (2011).

29. B. Günsel, C. Sezgin, J. Krajewski, SLEEPINESS DETECTION FROM SPEECH BY PERCEPTUAL FEATURES. *2013 Ieee Int Conf Acoust Speech Signal Process*, 788–792 (2013).

30. 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).

31. 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).

32. 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).

33. T. Chi, P. Ru, S. A. Shamma, Multiresolution spectrotemporal analysis of complex sounds. *J. Acoust. Soc. Am.* 118, 887–20 (2005).

34. 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).

35. 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).

36. N. Mesgarani, S. Shamma, SPEECH PROCESSING WITH A CORTICAL REPRESENTATION OF AUDIO. *2011 Ieee Int Conf Acoust Speech Signal Process Icassp* 1, 5872–5875 (2011).

37. E. Thoret, T. Andrillon, D. Léger, D. Pressnitzer, Probing machine-learning classifiers using noise, bubbles, and reverse correlation. *J Neurosci Meth* 362, 109297 (2021).

38. 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).

39. N. Ding, *et al.*, Temporal modulations in speech and music. *Neuroscience and Biobehavioral Reviews*, 1–7 (2017).

40. 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).

41. 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).

42. F. Gosselin, P. G. Schyns, RAP: a new framework for visual categorization. *Trends Cogn Sci* 6, 70–77 (2002).

43. K. L. Knutson, K. Spiegel, P. Penev, E. V. Cauter, The metabolic consequences of sleep deprivation. *Sleep Med Rev* 11, 163–178 (2007).

44. A. J. Krause, *et al.*, The sleep-deprived human brain. *Nat Rev Neurosci* 18, 404–418 (2017).

45. M. W. L. Chee, J. C. Tan, Lapsing when sleep deprived: Neural activation characteristics of resistant and vulnerable individuals. *Neuroimage* 51, 835–843 (2010).

46. 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).

47. T. Kaufmann, *et al.*, The brain functional connectome is robustly altered by lack of sleep. *Neuroimage* 127, 324–332 (2016).

48. 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).

49. C.-S. Hung, *et al.*, Local Experience-Dependent Changes in the Wake EEG after Prolonged Wakefulness. *Sleep* 36, 59–72 (2013).

50. 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).

51. B. Faraut, V. Bayon, D. Léger, Neuroendocrine, immune and oxidative stress in shift workers. *Sleep Med Rev* 17, 433–444 (2013).

52. F. Sauvet, *et al.*, Effect of acute sleep deprivation on vascular function in healthy subjects. *J Appl Physiol* 108, 68–75 (2010).

53. V. P. Martin, B. Arnaud, J.-L. Rouas, P. Philip, Does sleepiness influence reading pauses in hypersomniac patients? *Speech Prosody 2022*, 62–66 (2022).

54. 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).

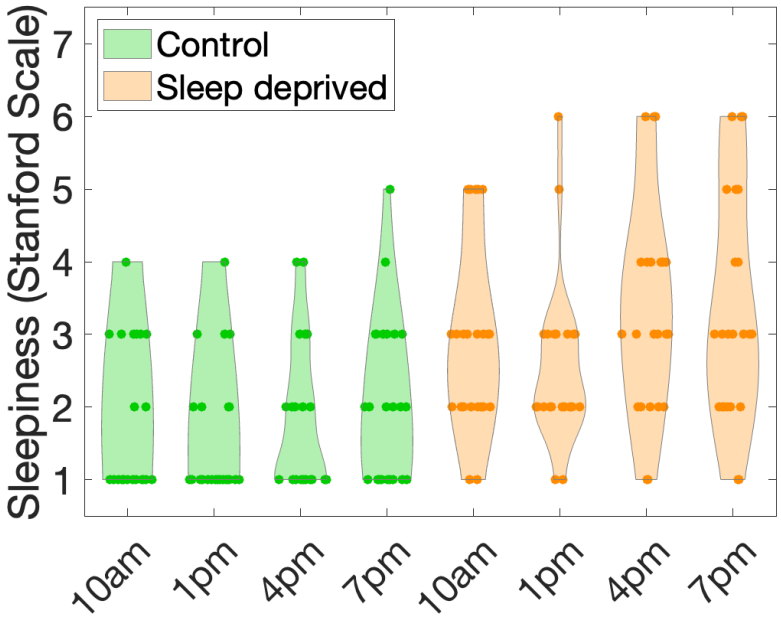
55. E. Thoret, P. Depalle, S. Mcadams, Perceptually Salient Regions of the Modulation Power Spectrum for Musical Instrument Identification. *Front. Psychol.* 8, 1–10 (2017).

56. B. N. Pasley, *et al.*, Reconstructing Speech from Human Auditory Cortex. *Plos Biol* 10, e1001251 (2012).

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

Figure 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 the data for individual participants.

A



B

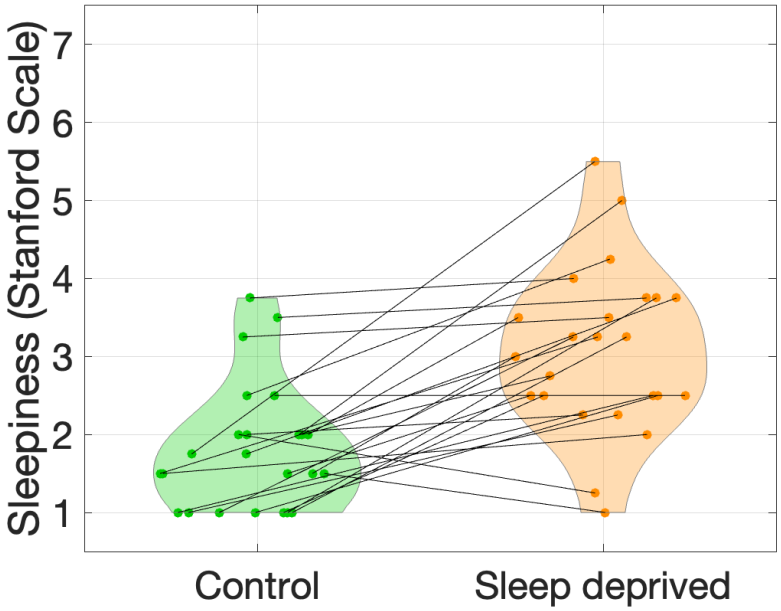


Figure 2. Acoustic analyses. **A.** Spectro-Temporal Modulations before sleep deprivation. Projections on the rate-scale and rate-frequency plane are shown. Arbitrary model units. **B.** As in A., but after sleep deprivation. **C.** Acoustic difference before and after sleep deprivation: 2 * abs(B-A) / (A+B). Units in percents.

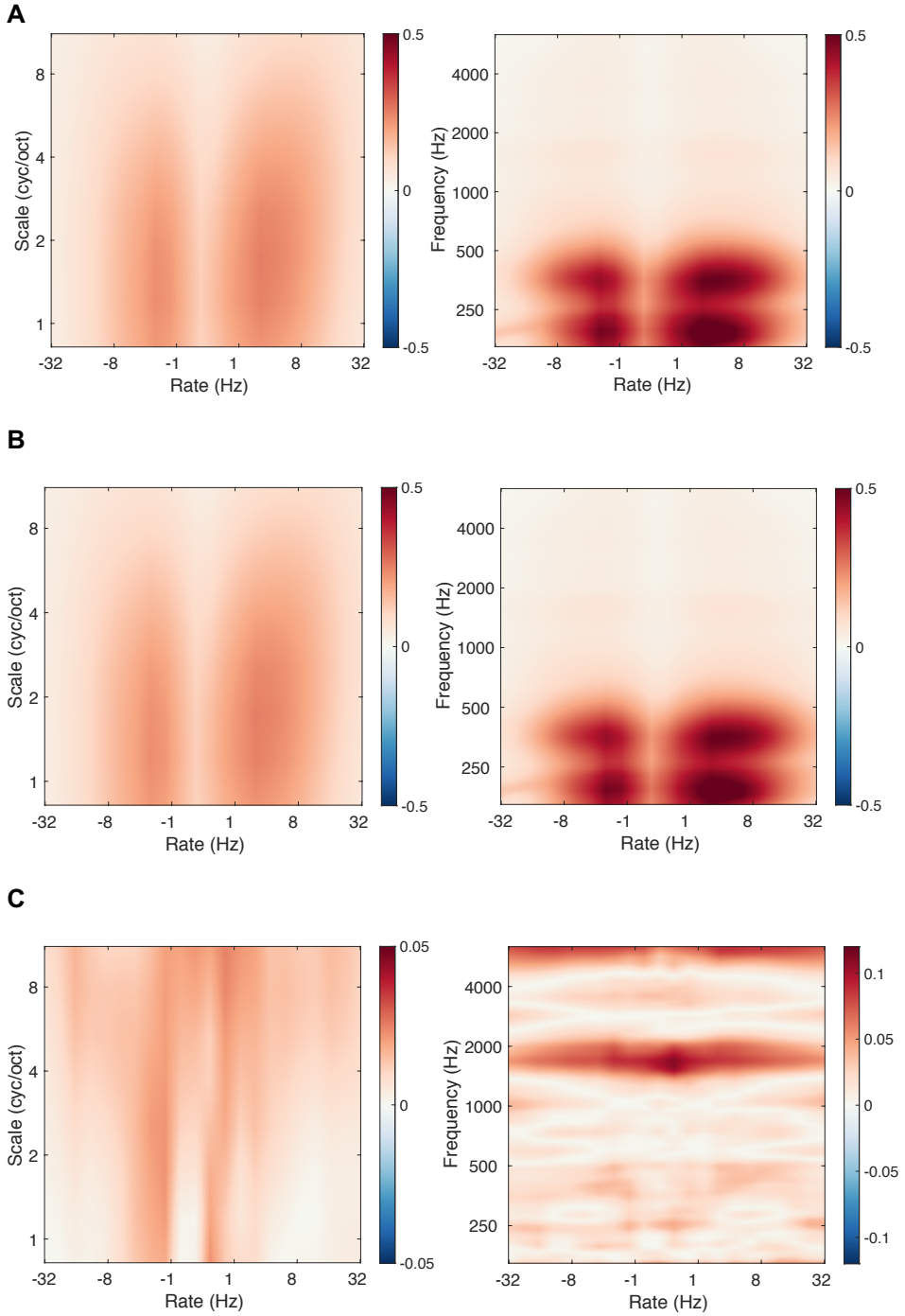


Figure 3. Machine learning classification results. Balanced Accuracies for the population-level classifier (red) and the participant-level classifiers (blue). Participants here and throughout the text are identified by their ranking # in classification accuracy. Error bars show the standard deviation of accuracies across 50 independent crossfold validations of each classifier.

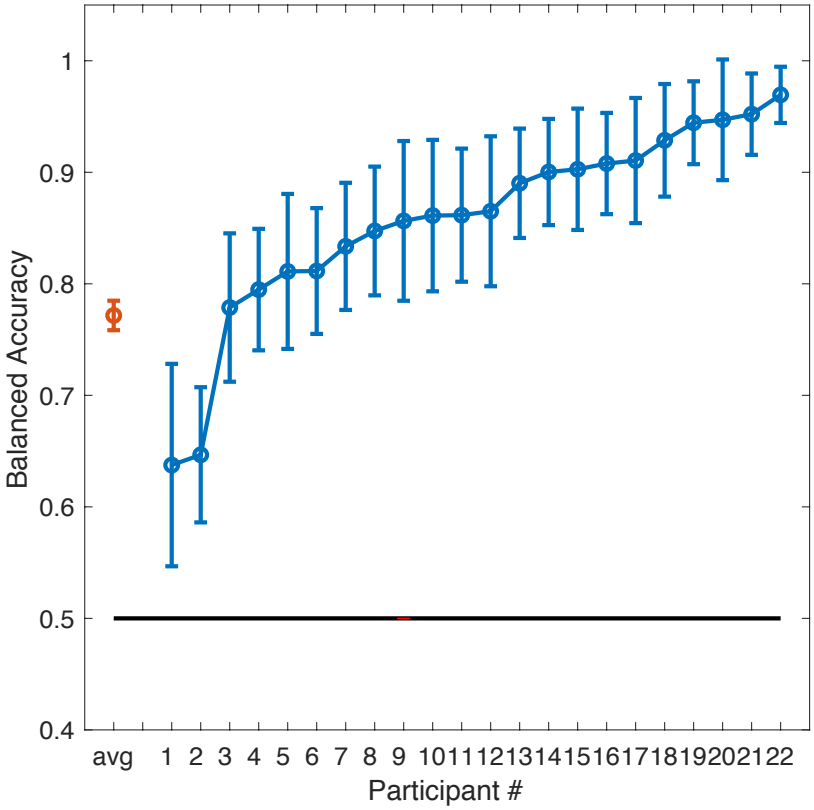


Figure 4. Interpretation of the population-level classifier. Discriminative features (see main text) are shown in the input STM space, for the rate-scale and frequency-scale projections. Red areas indicate features positively associated to sleep deprivation by the classifier. Blue areas correspond to 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 classifier.

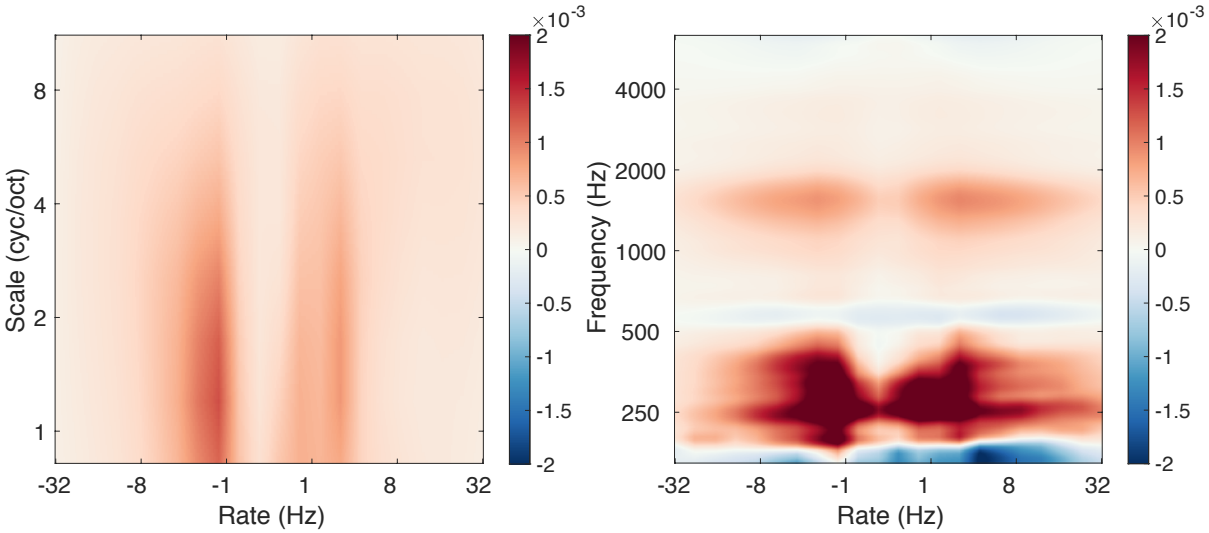


Figure 5. Interpretation of the participant-level classifiers. **A.** As for Figure 4, but for individual participants identified by their participant #. **B.** Projection of participants # in the interpretation-PCA space of all participant's masks (see text for details). **C.** Variance of the idealized masks, in correlation value, along the first two dimensions of the interpretation-PCA. Idealized masks are obtained by first sampling the PCA latent space between -2 and 2 for the two first dimesions with 30 values and then inverting the latent space into the input feature space by using the inverse transform of the PCA. Red areas show the discriminative features that vary the most along each interpretation-PCA dimension. Units: variance in the feature space.

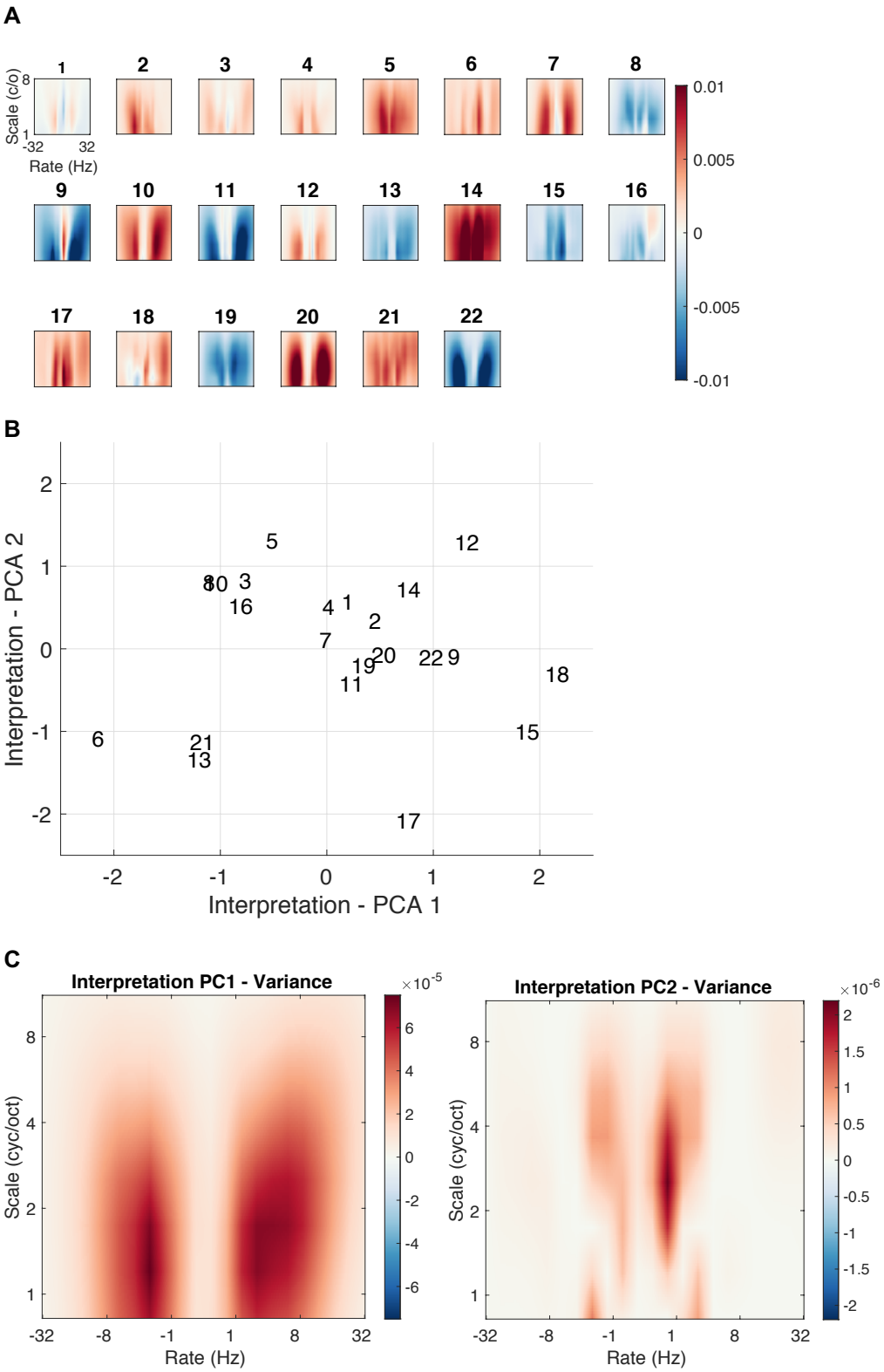
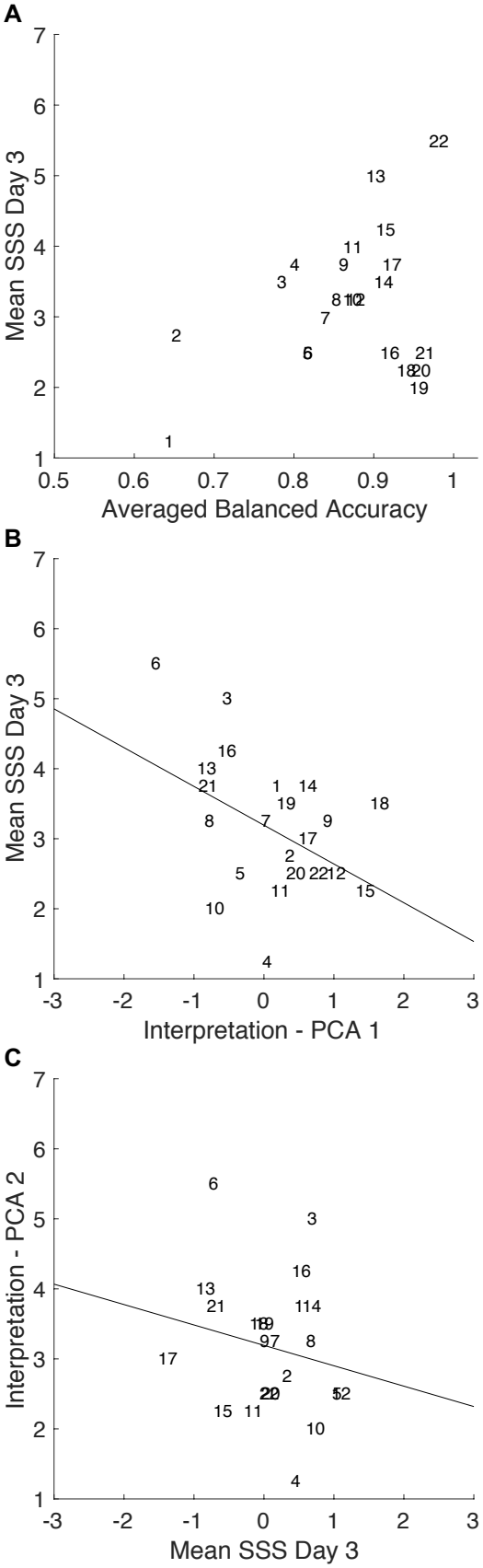



Figure 6. Relation between subjective sleepiness and voice classifiers. A. Subjective sleepiness is plotted as a function of balanced accuracy of each participant-level classifier. **B.** The coordinate of each participant-level classifier on the first dimension of the interpretation-PCA space is plotted as a function of subjective sleepiness. **C.** As in B., but for the second dimension of the interpretation-PCA space.