The Objective and Subjective Sleepiness Voice Corpora

Vincent P. Martin, Jean-Luc Rouas, Jean-Arthur Micoulaud-Franchi, Pierre Philip

▶ **To cite this version:**

Vincent P. Martin, Jean-Luc Rouas, Jean-Arthur Micoulaud-Franchi, Pierre Philip. The Objective and Subjective Sleepiness Voice Corpora. 12th Edition of its Language Resources and Evaluation Conference., May 2020, Marseille, France. pp.6525-6533. hal-02489433

HAL Id: hal-02489433
https://hal.science/hal-02489433

Submitted on 5 Mar 2020

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.

The Objective and Subjective Sleepiness Voice Corpora

Vincent P. Martin[1], Jean-Luc Rouas[1], Jean-Arthur Micoulaud Franchi[2], Pierre Philip[2]
[1]LaBRI - CNRS - UMR 5800 - Univ. Bordeaux - Bordeaux INP - F-33400 Talence (France)
[2]SANPSY - CNRS - USR 3413 - Univ. Bordeaux - CHU Pellegrin - F-33000 Bordeaux (France)
{vincent.martin, rouas}@labri.fr, {jean-arthur.micoulaud-franchi, pierre.philip}@u-bordeaux.fr

Abstract

Following patients with chronic sleep disorders involves multiple appointments between doctors and patients which often results in episodic follow-ups with unevenly spaced interviews. Speech technologies and virtual doctors can help improve this follow-up. However, there are still some challenges to overcome: sleepiness measurements are diverse and are not always correlated, and most past research focused on detecting instantaneous sleepiness levels of healthy sleep-deprived subjects. This article presents a large database to assess the sleepiness level of highly phenotyped patients that complain from excessive daytime sleepiness. Based on the Multiple Sleep Latency Test, it differs from existing databases by multiple aspects. First, it is composed of recordings from patients suffering from excessive daytime sleepiness instead of sleep deprived healthy subjects. Second, it incites the subjects to sleep contrary to existing stressing sleepiness deprivation experimental paradigms. Third, the sleepiness level of the patients is evaluated with different temporal granularities - long term sleepiness and short term sleepiness - and both objective and subjective sleepiness measures are collected. Finally, it relies on the recordings of 94 highly phenotyped patients, allowing to unravel the influences of different physical factors (age, sex, weight, ...) on voice.

Keywords: Sleepiness measurement and detection, Innovative healthcare, Read speech

1. Introduction

1.1. Motivations

One of the major challenges for diagnosing and treating neuro-psychiatric pathologies is symptom quantification and follow-up of chronic patients in order to adapt treatment and measure early relapses. Such an ecological monitoring is possible thanks to connected medical devices (measuring for instance weight, blood pressure or physical activities) but crucial information about how the patients report clinical symptoms like fatigue or sleepiness are difficult to measure. Regular in-person appointments between doctors and patients are useful but miss a large part of variability of symptoms at home in response to treatment. Furthermore, the growing number of patients increases the queuing time and often results in episodic follow-ups with unevenly spaced interviews.

Apart from the clinical interviews, it is nonetheless possible to measure some symptoms (*e.g.* sadness or sleepiness) with a range of behavioural analysis techniques: looking at eye movements and examining verbal expressions or body movements (Poursadeghiyan et al., 2018; Khan and Mansoor, 2008). Thanks to recent advances in speech processing, it seems now possible to detect precise cues in voice allowing to characterise the state of a speaker. This could potentially allow to measure the level of sleepiness, fatigue or sadness (Cummins et al., 2018). This method has multiple advantages as recording voice data is not invasive and it neither requires specific sensors nor complex calibration processes. It can thus be set up in various environments, outside laboratories, and allows regular and non-restrictive monitoring of patients.

It has already been shown that a virtual doctor using a semi structured interview as a diagnostic tool is well accepted by the patient (Philip et al., 2017). We wish to complete the analysis carried out using this method by recording the voice of the patients to determine their level of sleepiness. However, there are a few challenges to overcome to successfully reach that goal. Several subjective and objective methods have been designed to measure instantaneous sleepiness but they do not necessarily measure the same dimension of a common complaint (Shahid et al., 2011). Fatigue, depression and sleepiness can also be clinically misclassified unless a physician makes a clear investigation of the three previous symptoms which is quite difficult because of frequent co-morbidities. Finally, previous research on automatic detection only aimed at estimating instantaneous sleepiness for sleep-deprived healthy subjects. The most commonly used ones are presented and discussed in the next section.

1.2. Existing databases

The Sleepy Language Corpus (Schuller et al., 2011) - SLC - is the most used corpus on sleepiness detection through voice (Cummins et al., 2018; Martin et al., 2019). It consists of multiple speech tasks conducted in parallel of other sleeping-deprivation studies. The speakers are 99 German volunteers and the 9089 corresponding speech samples are mainly in German (or sometimes in English). The vocal tasks range from sustained vowels to spontaneous speech, including reading of novels or air traffic control commands. The sleepiness level of the speakers is labelled by the mean of three Karolinska Sleepiness Scale - KSS (Åkerstedt and Gillberg, 1990), one filled by the subject and two by trained external annotators. Setting the KSS limit between Sleepy (SL) and Non-Sleepy (NSL) samples to 7.5, the two classes are quite imbalanced (resp. 35% and 65% of the database). This imbalance probably comes from the experimental design that incites the subjects to maintain wakefulness: as they are in a stimulating environment (with driving tasks among others), subjects tend to stay awake and fight sleepiness. More information about the experimental setup of the

recordings can be found in (Krajewski et al., 2009; Golz et al., 2007). For extensive details about the dataset and the different experiments composing it, we invite the reader to see (Schuller et al., 2013).
Other databases are mostly used only by the authors who provided them. We will therefore not describe them in details but two examples are worth mentioning. First, a study conducted with 55 sleep-deprived American healthy subjects who were asked to freely answer open questions (McGlinchey et al., 2011). Their sleepiness level was self-evaluated using the Stanford Sleeping Scale (Hoddes et al., 1973). Another study recorded read speech from 22 French sleep-deprived subjects, evaluating their sleepiness using Electroencephalography (EEG) (Boyer et al., 2016).
Although very interesting results have been drawn from all these databases, leading to the possibility to assess the sleepiness level using voice recordings, most of the cited studies focused on analysing speech from sleep-deprived healthy subjects. Our goal is however quite different since we investigate patients suffering from sleepiness troubles which are related to long-term or chronic diseases. That is why none of the above-mentioned databases seem to be suited to our purposes, for the following reasons.
First, they rely on few subjects: except those using the SLC, only one study (McGlinchey et al., 2011) presents results over a quite large database (55 subjects). Since there are numerous factors that may influence speech while being unrelated to the sleepiness level (i.e. age, sex, weight, neck size, ...), an important number of speakers is needed to take into account all these parameters and the inherent variability of the general population.
Second, the sleepiness level is mostly evaluated using a single measurement, be it the KSS as in the SLC, the Stanford Sleepiness Scale (McGlinchey et al., 2011; Krajewski and Kroger, 2007), the Karolinska Drowsiness Test (Boyer et al., 2016) or Electroencephalography (EEG) (Golz et al., 2007; Dhupati et al., 2010; Boyer et al., 2016; Sparrow et al., 2019). Not only few studies collected both objective and subjective sleepiness measures, making complicated to establish their respective influence on voice, but the diversity of the scales curbs the comparison between the results.
Finally, all of these studies are based on sleep deprivation, that does not comply with ecological conditions. Moreover, this paradigm usually tends to create imbalances in the sleepiness level of the patients. Indeed, previous efforts of our team to estimate objective sleepiness from voice included the construction a corpus based on the Maintenance of Wakefulness Test (Mitler et al., 1982) - MWT - with 71 patients. Using this preliminary dataset did not allow us to carry on successful experiments on the determination of vocal biomarkers. As a matter of fact, this experimental design incited the patients fight against sleep, leading to a low percentage of sleepy patients (less than 15%). Moreover, a saturation effect appeared on 62% of the recordings because the subjects did not fall asleep at all.

1.3. Towards a new recording protocol

To study the link between voice and objective and subjective sleepiness on patients suffering from excessive daytime sleepiness, and to ensure having enough sleepy patients in this dataset and get more relaxing conditions during the recordings, we wish to apply a new experimental recording procedure based on the Multiple Sleep Latency Test-MSLT (Littner et al., 2005). Since this test encourages subjects to sleep, we expect them to start falling asleep more easily and thus have a more balanced dataset. Instead of investigating vocal biomarkers of the fight against sleep inducing the wakefulness system, our experimental design should highlight vocal clues involving the sleep onset system. The MSLT procedure does not induce stress, which could also modify the voice expression of the patients. In fact, they are recorded in a familiar environment (they arrive the day before the experiment) and the reading tasks are selected such as to not induce stress.
The database should provide sleepiness measurements at different time granularities, with objective and subjective measures. On the one hand, it should give long term measurements, using self-reported questionnaires on the sleepiness habits of the patients. On the other hand, it should give instantaneous measures of sleepiness such as Polysomnographic measures (the objective MSLT measure), or the results of questionnaires (the KSS or the Cartoon Faces Questionnaire (Maldonado et al., 2004) for example). Finally, the patients should be highly phenotyped, allowing to unravel the influence of the individual factors on voice and to use subgroups of sleep disorders if necessary.
To answer these questions, we describe in this paper the details of the creation of our new databases and provide some insights pertaining to its additional contents (medical data, etc.)
This paper is thus structured as follows. In Section 2. we provide a description of the MSLT database and the MSLT procedure. In Section 3. we present and justify our choice for the texts used in the reading tasks. Section 4. provides an overview of the database and discussion about the obtained results is made in Section 5. Finally, conclusions and future work are presented in Section 6.

2. Presentation of the database

2.1. Description of the MSLT database

The MSLT database has been elaborated and is collected at the Bordeaux University Hospital Sleep Clinic, France. All the recorded patients suffer from of excessive daytime sleepiness or nocturnal breathing disorders. A summary of the database is presented in Table 2.

2.1.1. Procedure of the MSLT

The procedure of the MSLT is the following. The patients are welcomed the evening prior to the exam for a first night of polysomnography. The day of the exam, they are asked to take a nap every two hours at 9am, 11am, 1pm, 3pm and 5pm. Approximately ten minutes before the beginning of the exam, the voice of the patients is recorded and they fill the Karolinska Sleepiness Scale Questionnaire - KSS (Åkerstedt and Gillberg, 1990). After completing the Cartoon Faces scale, lights are switched off and the test begins. The patients have a 20 minutes period to fall asleep: if they stay awake during this period, the test is terminated. If they fall asleep, the recording is extended for a 15 minutes period. After that the lights are turned off, one epoch of

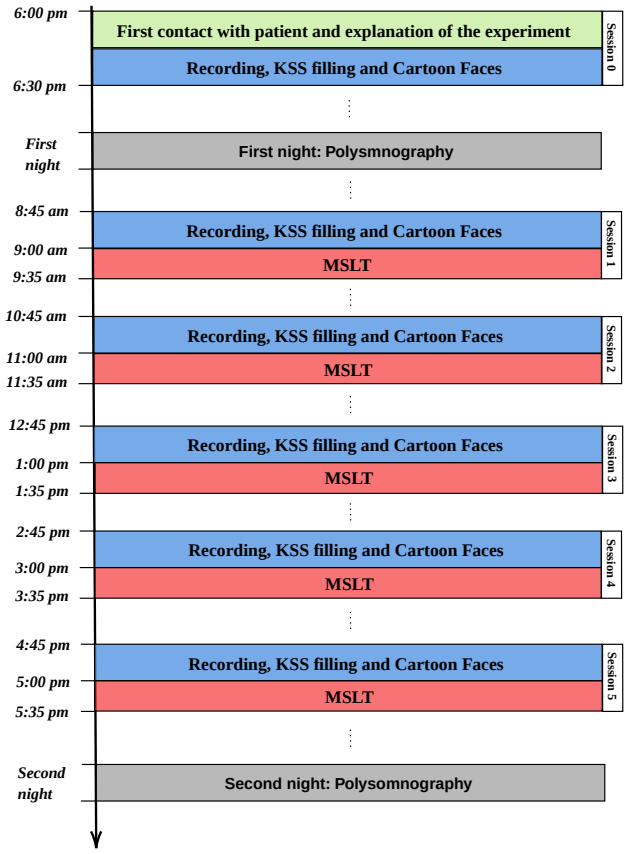

Figure 1: Typical time table of a patient during the recording of the MSLT database.

any sleep stage is required to define sleep onset (Littner et al., 2005). The maximum length of the sleep onset period being 20 minutes, all the MSLT values are under or equal to 20 minutes. This procedure is summarised in the Figure 1.

2.1.2. Recording procedure

Each patient reads six different texts that are the same at constant session. These texts are presented in Section 3. The first text is read during the reference recording (Session 0, see Figure 1), carried out the day before the exam around 6pm, time at which the circadian cycle is at its apex (Sedgwick, 1998). This recording allows the patient to familiarise with the procedure and the material, after having being informed of all the details of the experiment and signed a consent form. The recording procedure is the following. First, the patients are asked to quietly read the text, to acquaint themselves with the content of the text. Second, the patients fill the KSS questionnaire. Third, they are asked to read the text aloud and their voice is recorded. This procedure is the same for each iteration of the test. All the recordings are made in the room in which the patient takes the test, with an omni-directional Audio-technica AT4022 microphone connected to a Tascam DR-100 MKIII audio recorder. To ensure minimum alteration of the recordings due to environment and position of the vocal apparatus, the patients are either in their bed or installed at their desk, the positions of the patient and the microphone being the same for all the iterations.

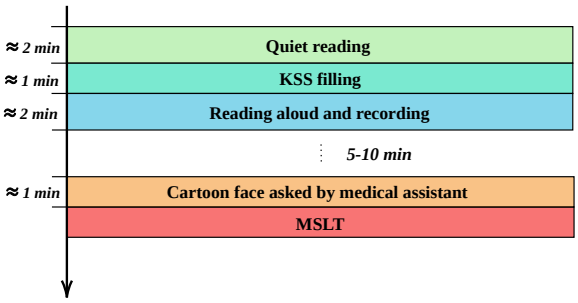

Figure 2: Detail of the procedure to record the voice of patients before a MSLT iteration

This procedure is summarised in Figure 2. This leads to a total of 12h 6min and 1s of audio recordings in our corpus (resp. 10h 18min 2s without the session 0).

3. Read texts

This section presents the choice of the tasks chosen to record voice of the subjects and justify the choice of the texts.

3.1. Reading tasks

As our subjects are patients, they could have untreated hypersomnia. This could lead to difficulties to carry out tasks involving a high cognitive load. As reading has a lower cognitive load than spontaneous speech (Christodoulides, 2016), we choose to focus on reading tasks. Furthermore, such a task assures valid comparison between patients since all the patients are asked to read the same texts. The recordings should also be less contaminated by emotions compared with spontaneous speech.

3.2. Choice of the texts

The texts have to be as neutral as possible regarding the emotional state of the patients (neither boring nor too exiting) to avoid an alteration of their sleepiness state. This constraint is completed by the need of simple grammar and vocabulary, allowing readers with different reading skills. As it is already widely used in phonetic studies (Raake, 2002; Goldman et al., 2016), we choose extracts from *Le Petit Prince* by Antoine de Saint-Exupéry. The first chapters of the French version are truncated so as to be approximately 200 words long while keeping the coherence of the meaning of the text. We thus extracted six texts corresponding to the six iterations presented in Section 2.1. The texts are printed with the *Times New Roman* font with a 15pt size to ensure a good readability by all patients. Using this protocol, the length of the recordings varies between 50 seconds and 2 minutes, depending on the reading capacities of the subject (mean duration: 80.8 seconds, std: 21.5 seconds).

3.3. Reading level of the patients in the MSLT database

To take into account the reading skills of the patients and the difficulties of the texts, the ELFE score (*Évaluation de*

la Lecture en FluencE - Evaluation of the reading with fluency) developed in (Cogniscience, 2008) is measured for each reading. It consists on subtracting the number of mistakes not handled by the patient to the number of words correctly read in one minute. The mean ELFE score of all the patients depending on the text and the moment of the day is represented in Figure 3. For each iteration, the reference of the session, the hour, the number of words and the difficulty level of the texts (as estimated by the ratio of the number of words containing more than two syllables over the total number of words) are specified on the label of the horizontal axis. At first glance, we observe that the variations of the ELFE score across the session is influenced by the sex and the subjective sleepiness state (KSS). These variations could however also be influenced by the texts, which may not have the same difficulty level. To study these different influences, we conduct a multivariate ANOVA taking into account the KSS, the difficulty of each text and the sex of the subjects to explain the variations of the ELFE score. The influence of the session is dominant ($F = 49.0$, $p < 10^{-16}$), but the KSS ($F = 2.4$, $p < 10^{-2}$) and the sex (cross-interaction between sex and KSS, $F = 2.4$, $p < 5 \times 10^{-2}$) also have an influence on the ELFE score. We assume that the major differences between the ELFE distributions across the experience are only due to variations of subjective sleepiness and sex, the minor differences in difficulty level having a negligible effect.
Furthermore, a Spearman's ρ led to the conclusion that the mean ELFE score is directly correlated to the social level of the reader ($\rho = 0.34$, $p = 0.0008$). As women recorded in this database have a higher social level than men (see last line of Table 2), the influence of the sex factor over the ELFE score may be explained by a difference of social level.

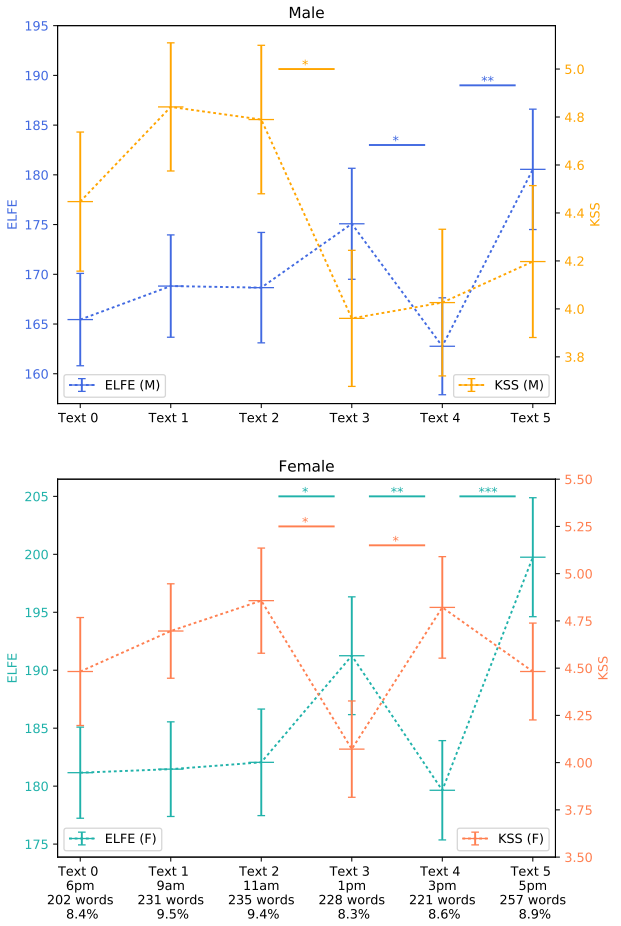

Figure 3: Distribution of the ELFE score and the KSS by hour plotted with Standard Error of Mean. X-axis labels: first row: text; second row: hour; third row: total number of words in the text; fourth row: ratio of difficult words in the text.
Mann-Whitney tests (*: $p < 5 \times 10^{-2}$, **: $p < 10^{-2}$, ***: $p < 10^{-3}$, ****: $p < 10^{-4}$)

4. Medical data

Sleepiness estimation faces two main challenges. On the one hand, as most our subjects are patients with Excessive Daytime Sleepiness, their objective sleepiness measured by EEG does not necessarily correlates with their perceptual sleepiness. On the other hand, the temporal granularity of the sleepiness estimation varies from one questionnaire to another. To study the different parameters influencing voice production, the database includes both subjective and objective measures, at two different time levels: few minutes before the MSLT iteration (designated as *MSLT iteration scale*) and the habits of the patients on several days or weeks before the test (designated as *Patients scale*). As this database has been elaborated in France, all the questionnaires mentioned in this article are in French. A summary of the database is proposed in Table 2.

4.1. MSLT and subjective sleepiness scales

During the MSLT test, three sleepiness measures are collected, then averaged over the five iterations (Session 1-5) of the protocol explained in Section 2.1.

4.1.1. Subjective sleepiness scales

Two perceptual questionnaires are filled by the patient during the interview before each MSLT iteration: the KSS and the Cartoon Faces. The KSS is a nine items questionnaire going from 1: 'extremely alert' to 9: 'Very sleepy, great effort to keep awake, fighting sleep' (resp. 'Très éveillé' and 'Très somnolent, avec de grands efforts pour rester éveillé, luttant contre le sommeil' in French). It is the most used sleepiness questionnaire in studies about influence of sleepiness on speech (Schuller et al., 2019; Schuller et al., 2011) and has already be proved a confident measure of subjective sleepiness (Åkerstedt et al., 2014).
The Cartoon Faces Sleepiness Scale consists on five cartoon faces reflecting five different states of sleepiness. It has the advantage of not necessitating the comprehension of any language and is easier and more intuitive to answer when, for example, dealing with patient having severe sleep disease.

4.1.2. Clinical Data (MSLT scale)

These two subjective measures are completed by the objective measure of EEG during the iteration of the MSLT, providing the time needed by the patient to fall asleep after the beginning of the test. In the following, this measure will be denominated 'MSLT iteration value'.

Category of questionnaire	Questionnaires	Reference	Description
Objective sleepiness measure			
Sleepiness	MSLT value	(Littner et al., 2005)	Time (in min) between beginning of the test and sleeping onset (0-20 min)
Subjective sleepiness measures (MSLT iteration scale)			
Sleepiness	KSS	(Åkerstedt and Gillberg, 1990)	9 items about sleepiness (1-9)
	Cartoon Faces	(Maldonado et al., 2004)	5 graphical items about sleepiness (0-4)
Subjective sleepiness and co-morbidity factors measures (Patient scale)			
Sleepiness Fatigue	Epworth Sleepiness Scale (ESS)	(Shahid et al., 2011, p.149)	8 items about daytime sleepiness (0-24)
	Insomnia Severity Index (ISI)	(Shahid et al., 2011, p.191)	7 items about insomnia (0-28)
	Functional Outcomes of Sleep Questionnaire-10 (FOSQ-10)	(Shahid et al., 2011, p.179)	10 items about the impact of daytime sleepiness on activities of daily living (10-40)
	Fatigue Severity Scale (FSS)	(Shahid et al., 2011, p.167)	9 items about fatigue (9-63)
	Toronto & Hospital Alertness Test	(Shahid et al., 2011, p.391)	10 items to measure alertness (0-50)
	Part A of ADHD Self-Report Scale (ASRS)	(Schweitzer et al., 2001)	6 items about attention-deficit/hyperactivity disorder (0-24)
	Barcelona Scale	(Guaita et al., 2015)	2 items about sleepiness (0-6)
	Hobson Scale	(Hobson et al., 2002)	4 items about excessive daytime sleepiness (0-16)
Anxiety and Depression	Hospital Anxiety and Depression scale	(Zigmond and Snaith, 1983)	7 items about depression 7 items about anxiety (0-21)
Alcohol	Cut-down, Annoyed, Guilty, Eye-opener Questionnaire (CAGE)	(Shahid et al., 2011, p.415)	4 items about alcohol consumption (0-4)
Cigarettes	Cigarette Dependence Scale, short version (CDS-5)	(Courvoisier and Etter, 2008)	5 items about cigarettes dependence (5-25)
Social level measures			
Reading Level	Évaluation de la Lecture en FluencE (ELFE)	(Cogniscience, 2008)	Number of words read in one minute minus the number of errors
Education level	-	-	Years of study after the French Certificate of general education

Table 1: Medical information (patient scale and MSLT iteration scale) collected for the database.

4.2. Medical questionnaires (Patient scale)

Multiple questionnaires and medical measures are collected about the patient to take into account two aspects of the challenge. On the one hand, all the physiological parameters that could affect the vocal production are measured and integrated to the database. On the other hand, medical questionnaires that allow the estimation of the different components of sleepiness, fatigue and depression are collected. These clinical measures are completed with a Polysomnography the night preceding the exam, the collect of the pathologies and the treatments that can affect voice (psychostimulants, myorelaxants, ...) and diverse physiological measures such as height, weight, age, neck size, ... The clinical data collected anonymously in this study are presented in Table 1.

5. Discussion

The physiological characteristics of the patients and their answers to the questionnaires are discussed below. In the following, we distinguish between sleepy and non-sleepy subjects according to the mean value of their MSLT iterations. The typical value used by clinicians to make this

Questionnaires	Mean MSLT ≤ 8 (SL)			Mean MSLT > 8 (NSL)			All		
	M	F	Both	M	F	Both	M	F	Both
Objective sleepiness measure									
mean MSLT (0-20)	5.0 (1.9) ****	4.5 (2.2) ****	4.8 (2.0) ****	12.9 (3.7) ****	13.6 (3.2) ****	13.4 (3.4) ****	9.8 (4.9) **	12.3 (4.4) **	11.3 (4.8)
Physiological measures									
Number	15	8	23	23	48	71	38	56	94
Age	37.8 (18.2)	29.9 (6.7)	35.0 (15.7)	37.8 (15.1)	36.1 (12.1)	36.6 (13.2)	37.8 (16.4)	35.2 (11.7)	36.3 (13.9)
Body Mass Index	25.7 (4.1)	25.6 (5.7)	25.6 (4.7) *	26.6 (4.8) ***	23.3 (6.0) ***	24.4 (5.8) *	26.2 (4.5) ***	23.6 (6.0) ***	24.7 (5.6)
Height (m)	1.76 (0.05) ***	1.64 (0.06) ***	1.72 (0.08)	1.78 (0.05) ****	1.64 (0.06) ****	1.69 (0.09)	1.77 (0.05) ****	1.64 (0.06) ****	1.69 (0.09)
Weight (kg)	79.2 (12.7)	69.3 (17.9)	75.8 (15.4) *	84.3 (14.7) ****	62.4 (13.5) ****	69.5 (17.3) *	82.3 (14.2) ****	63.4 (14.4) ****	71.0 (17.1)
Neck size (cm)	41.8 (3.2) **	36.1 (3.1) **	39.8 (4.2) *	41.8 (3.2) ****	35.6 (3.4) ****	37.6 (4.5) *	41.8 (3.2) ****	35.6 (3.4) ****	38.1 (4.5)
Cigarettes/day	1.5 (3.4)	2.8 (7.3)	2.0 (5.1)	1.5 (4.0)	2.3 (6.3)	2.0 (5.7)	1.5 (3.8)	2.4 (6.5)	2.0 (5.6)
Alcohol glasses/day	0.2 (0.5) *	0.0 (0.1)	0.1 (0.4)	0.6 (1.1) *	0.1 (0.3)	0.3 (0.7)	0.4 (0.9) *	0.1 (0.3) *	0.2 (0.6)
Subjective sleepiness measures (MSLT iteration scale)									
Mean KSS (1-9)	4.0 (1.1) *	5.2 (1.1) *	4.4 (1.2)	4.6 (1.5)	4.5 (1.3)	4.5 (1.3)	4.4 (1.4)	4.6 (1.3)	4.5 (1.3)
Mean Cartoon Faces (0-5)	1.5 (0.6)	1.8 (0.5)	1.6 (0.6)	1.6 (0.6)	1.7 (0.6)	1.6 (0.6)	1.6 (0.6)	1.7 (0.6)	1.6 (0.6)
Subjective sleepiness and co-morbidity factors measures (Patient scale)									
Fatigue	11	7	18	19	45	64	30	52	82
Snoring	3	3	6	7	10	17	10	13	23
Hypertension	3	2	5	5	2	7	8	4	12
Observed Sleepiness Apnea	4	2	6	6	7	13	10	9	19
ESS (0-24)	14.9 (5.1)	17.9 (4.6)	16.0 (5.1)	12.8 (6.0)	14.8 (4.4)	14.1 (5.0)	13.6 (5.7)	15.2 (4.5)	14.6 (5.1)
ISI (0-28)	13.3 (5.3)	14.4 (5.5)	13.7 (5.4)	16.0 (5.3)	15.1 (5.3)	15.4 (5.3)	14.9 (5.5)	15.0 (5.3)	14.9 (5.4)
FOSQ-10 (10-40)	25.1 (7.5)	20.0 (8.3)	23.3 (8.1)	21.7 (5.3)	21.3 (7.5)	21.4 (6.8)	23.0 (6.5)	21.1 (7.6)	21.9 (7.2)
FSS (9-63)	35.0 (10.7) ***	49.0 (10.6)	39.9 (12.6) ***	49.7 (10.4) ***	49.3 (11.2)	49.4 (11.0) ***	43.9 (12.8) *	49.3 (11.1) *	47.1 (12.1)
Toronto (0-50)	28.2 (8.8) **	24.7 (7.5)	27.0 (8.5) **	21.7 (6.5) **	22.9 (8.2)	22.5 (7.7) **	24.3 (8.1)	23.1 (8.1)	23.6 (8.2)
ASRS (0-24)	10.9 (5.7)	12.1 (4.5)	11.3 (5.3)	13.9 (5.2)	11.8 (4.9)	12.5 (5.1)	12.7 (5.6)	11.8 (4.8)	12.2 (5.2)
Barcelona (0-6)	2.4 (1.3)	2.4 (1.3)	2.4 (1.3)	2.2 (0.9)	2.3 (0.9)	2.3 (0.9)	2.3 (1.1)	2.3 (1.0)	2.3 (1.0)
Hobson (0-12)	4.3 (2.4)	5.9 (3.2)	4.8 (2.8)	3.8 (2.5)	4.1 (2.3)	4.0 (2.4)	4.0 (2.4)	4.3 (2.6)	4.2 (2.5)
HAD Depression (0-21)	4.5 (3.2) **	4.8 (4.4)	4.6 (3.7) **	7.1 (2.7) **	7.0 (4.6)	7.0 (4.1) **	6.1 (3.2)	6.7 (4.6)	6.4 (4.1)
HAD Anxiety (0-21)	6.3 (3.0) *	8.0 (3.4)	6.9 (3.2)	9.0 (4.6) *	8.3 (4.0)	8.5 (4.2)	8.0 (4.3)	8.2 (3.9)	8.1 (4.1)
CAGE 0-4	0.3 (1.0) *	0.2 (0.4)	0.3 (0.8)	0.7 (0.9) *	0.2 (0.6)	0.4 (0.7)	0.5 (0.9)	0.2 (0.6)	0.4 (0.8)
CDS-5 5-25	6.5 (3.4)	7.1 (5.6)	6.7 (4.3)	6.6 (3.5)	7.3 (4.9)	7.0 (4.5)	6.5 (3.5)	7.2 (5.0)	7.0 (4.4)
Social level measures									
Mean ELFE	176.2 (31.0)	176.2 (41.5)	176.2 (35.0)	167.9 (32.8) **	188.6 (31.8) **	181.9 (33.6)	171.2 (32.4) **	186.8 (33.6) **	180.5 (34.0)
Education level	3.9 (2.2)	4.8 (1.4)	4.2 (2.0) *	4.5 (2.3) *	5.7 (2.6) *	5.3 (2.6) *	4.3 (2.2) **	5.5 (2.5) **	5.0 (2.5)

Table 2: Summary of the data collected for the database. The different colors represent the result of Mann-Whitney tests. Green: Sig. Difference between sex ind. from the sleepiness level. Red: Sig. Difference between sleepiness group ind. from the sex. Blue: (resp. Orange) Difference between Sleepy and Non-Sleepy men (resp. women).
(*: $p < 5 \times 10^{-2}$, **: $p < 10^{-2}$, ***: $p < 10^{-3}$, ****: $p < 10^{-4}$)

distinction is set to 8 minutes. The responses to questionnaires are thus studied from two perspectives: the influence of the MSLT group (mean MSLT iteration values $\leq$ or $>$ 8), the influence of sex, and their cross influence.

5.1. Population characteristics

Despite the application of the MSLT protocol, we still have fewer sleepy subjects than non-sleepy ones (23 vs. 71). Since the recordings are on-going we will focus on recording patients that could allow to have a more balanced database. As patients are however highly phenotyped, we can still observe differences between the two groups (SL and NSL).

The average age of our patients is 36.3 years with a standard deviation of 13.9. Our sleepy patients are slightly younger than their non-sleepy counterparts (35.0 vs. 36.6) which is due to a younger female sleepy population (29.9 average).

Sleepy patients are globally a little heavier (mean weight 75.8kg vs. 69.5kg) and have a higher neck size (39.8cm vs. 37.6cm) and Body Mass Index (25.6kg.m^{-2} vs. 24.4kg.m^{-2}) than their non-sleepy counterparts. Even if these differences are statistically significant, they are small and should not create biases. There are also differences of height, weight, BMI and neck size between men and women, that already exist in the natural population in France (Verdot et al., 2013). The same difference exists concerning the alcohol consumption (Richard et al., 2019), small but significant difference (0.4 glasses per day) is observed between sleepy and non-sleepy men.

5.2. Subjective sleepiness (Iteration Scale)

Regarding the KSS, only a slight difference between men (4.0) and women (5.2) is observed for the sleepy subjects whereas a difference between Sleepy and Non-Sleepy distribution was expected. This could be due to a procedural fault. Indeed, in (Horne and Burley, 2010), physicians indicate that a five minutes settling down period is necessary and sufficient to allow participants to be in correct conditions to self-assess accurately their subjective sleepiness. In that case, it correlates with EEG (objective) measures. We assume that in the MSLT database, the patients do not have the time to fully relax and be in these conditions. As matter of fact, the KSS is filled between the two readings of the texts (the first being quiet, the second being at loud and recorded), i.e. less than two minutes after a change in the activity of the subject. Moreover, the patients are under time pressure when filling the questionnaire: they fill the KSS under the supervision of the research assistant who records the voice. This haste tends to not let patients take the time they need to accurately self-evaluate their subjective sleepiness.

The same observations is made for the Cartoon Sleepiness Scale: no differences is observed between the groups (sex or level of sleepiness). The value being asked to the patients less than five minutes after that they have been installed in their bed, they do not have the time to self-evaluate accurately their subjective sleepiness level.

Since our subjects are however patients suffering from excessive daytime sleepiness, our results are hardly comparable to those of (Horne and Burley, 2010) who studied healthy subjects.

Peculiar attention should be payed to this crucial but easy-to-set up condition for the subjective sleepiness measures at the iteration scale, which will confirm or infirm the correlation between KSS and objective sleepiness on our database. In addition all the patients are very sleepy and we might not have enough patients to observe differences in subjective measures which can be affected by many factors (age, weight, ...).

5.3. Subjective sleepiness and co-morbidity factors measures (Patient Scale)

Almost all the subjects (82 over 94) declare to feel tired (Fatigue item corresponding to the answer of the following question: "Do you feel tired, exhausted or sleepy during daytime ?") while only few are snorers or have already been seen making obstructive sleepiness apnea.

The ESS score is not different for sleepy (16.0) or non-sleepy (14.1) patients but all patients report very high level of sleepiness with a trend for higher ESS scores in the SL category. A larger sample size may have shown positive results (Sangal, 1999).

The sleepy and non-sleepy patients also have the same levels of insomnia (ISI 13.7 vs. 15.4), impact on everyday life (FOSQ-10, Hobson, Barcelona) and attention deficit/hyperactivity disorder (ASRS 11.3 vs. 12.5).

A counter-intuitive observation is that sleepy men have lower FSS (35.0) than their non-sleepy counterparts (49.7), meaning that the latter feel more exhausted. Moreover, the Toronto score is lower for non-sleepy men (28.2) than for the sleepy ones (21.7). The Toronto questionnaire is an inverted scale: it measures severe symptoms when near zero. As both the FSS and the Toronto scores have been shown to be influenced by anxiety and depression (Shahid et al., 2011) and the scores to the HAD (Depression and Anxiety questionnaire) are significantly higher for the non-sleepy men (7.1 and 9.0 resp.) than for sleepy men (4.5 and 6.3 resp.), the Toronto and FSS scores differences could be explained by depression and anxiety instead of differences in sleepiness or fatigue.

Finally, sleepy men have lower alcohol dependence scores (CAGE 0.3) than their non-sleepy counterparts (CAGE 0.7), which is consistent with their respective consumption. No difference is shown concerning the cigarettes addiction.

6. Conclusion & Perspectives

To achieve our goal of following patients suffering from chronic sleep disorders, we have introduced a new database for the detection of sleepiness through voice, with an experimental set up promoting several sleepiness measurements through subjective and objective components. These measures are given for different time granularities, allowing to refine the link between voice and the different types of sleepiness (long term vs. short term, subjective vs. objective). The substitution of the usual sleep deprivation experimental set-up by the MSLT procedure to have a more balanced dataset does not seem to have reach its full potential. Further recordings selecting the patients to balance the dataset will fix the imbalance between sleepy and non-sleepy subjects.

However, there are still differences observed between the two groups (SL and NSL) over the numerous phenotypical items collected from the patients. Besides some weakly significant physical differences between the two groups, this database allowed to highlight that subjective and objective sleepiness (KSS vs. MSLT and Cartoon faces vs. MSLT) do not correlate when dealing with patients suffering from excessive daytime sleepiness. Moreover, we observed that Toronto and FSS scores observed in our study are polluted by depression (HAD-depression): this illustrates the joint influence of fatigue and depression over sleepiness questionnaires. Furthermore, no difference has been observed for numerous questionnaires (ISI, FOSQ-10, Hobson, Barcelona). Special attention over this topic will be paid when using questionnaires to assess sleepiness.
Further work on this database will include the recording of targeted subjects to balance the dataset especially concerning sleepy women. We will also study thoroughly the different measurements used to assess the sleepiness level (both objective and subjective) and their correlation. Our next objective will be to elaborate a voice analysis system allowing to estimate sleepiness level from voice samples. Finally, such a system will be implemented within the virtual physician.

7. Acknowledgements

This work is carried out in the framework of the IS-OSA project funded by the French Region Nouvelle Aquitaine and by the SOMVOICE project sponsored by the Labex BRAIN (University of Bordeaux, France).

8. Bibliographical References

Boyer, S., El-Yagoubi, R., Tiberge, M., Ruiz, R., and Daurat, A. (2016). Paramètres Acoustiques de la Voix et Privation de Sommeil. In *CFA/VISHNO*.

Christodoulides, G. (2016). *Effects of Cognitive Load on Speech Production and Perception*. Ph.D. thesis, Université Catholique de Louvain.

Cogniscience. (2008). E.L.FE - Évaluation de la Lecture en FluencE. Technical report.

Courvoisier, D. and Etter, J.-F. (2008). Using item response theory to study the convergent and discriminant validity of three questionnaires measuring cigarette dependence. *Psychology of Addictive Behaviors*, 22(3):391–401.

Cummins, N., Baird, A., and Schuller, B. (2018). Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Health Informatics and Translational Data Analytics*, 151:1–54.

Dhupati, L. S., Kar, S., Rajaguru, A., and Routray, A. (2010). A novel drowiness detection scheme based on speech analysis with validation using simultaneous EEG recordings. In *IEEE - Int. CASE*, pages 917–921.

Goldman, J.-P., Honnet, P.-E., Clark, R., Garner, P. N., Ivanova, M., Lazaridis, A., Liang, H., Macedo, T., Pfister, B., Ribeiro, M. S., Wehrli, E., and Yamagishi, J. (2016). The SIWIS Database: A Multilingual Speech Database with Acted Emphasis. In *Interspeech*, pages 1532–1535.

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

Guaita, M., Salamero, M., Vilaseca, I., Iranzo, A., Montserrat, J. M., Gaig, C., Embid, C., Romero, M., Serradell, M., León, C., de Pablo, J., and Santamaria, J. (2015). The Barcelona Sleepiness Index: A New Instrument to Assess Excessive Daytime Sleepiness in Sleep Disordered Breathing. *Journal of clinical sleep medicine*, 11(11):1289–1298.

Hobson, D. E., Lang, A. E., Martin, W. R. W., Razmy, A., Rivest, J., and Fleming, J. (2002). Excessive daytime sleepiness and sudden-onset sleep in Parkinson disease: a survey by the Canadian Movement Disorders Group. *JAMA*, 287(4):455–463.

Hoddes, E., Zarcone, V., Smythe, H., Phillips, R., and Dement, W. C. (1973). Quantification of sleepiness: a new approach. *Psychophysiology*, 10(4):431–436.

Horne, J. and Burley, C. (2010). We know when we are sleepy: Subjective versus objective measurements of moderate sleepiness in healthy adults. *Biological Psychology*, 83(3):266–268.

Khan, M. I. and Mansoor, A. B. (2008). Real Time Eyes Tracking and Classification for Driver Fatigue Detection. In *Image Analysis and Recognition*, pages 729–738.

Krajewski, J. and Kroger, B. (2007). Using prosodic and spectral characteristics for sleepiness detection. In *Interspeech*, pages 1841–1845.

Krajewski, J., Batliner, A., and Golz, M. (2009). Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach. *Behavior Research Methods*, 41(3):795–804.

Littner, M. R., Kushida, C., Wise, M., Davila, D. G., Morgenthaler, T., Lee-Chiong, T., Hirshkowitz, M., Loube, D. L., Bailey, D., Berry, R. B., Kapen, S., and Kramer, M. (2005). Practice Parameters for Clinical Use of the Multiple Sleep Latency Test and the Maintenance of Wakefulness Test. *Sleep*, 28(1):113–121.

Maldonado, C. C., Bentley, A. J., and Mitchell, D. (2004). A Pictorial Sleepiness Scale Based on Cartoon Faces. *Sleep*, 27(3):541–548.

Martin, V. P., Rouas, J.-L., Thivel, P., and Krajewski, J. (2019). Sleepiness detection on read speech using simple features. In *10th Conference on Speech Technology and Human-Computer Dialogue*.

McGlinchey, E. L., Talbot, L. S., Chang, K.-h., Kaplan, K. A., Dahl, R. E., and Harvey, A. G. (2011). The Effect of Sleep Deprivation on Vocal Expression of Emotion in Adolescents and Adults. *Sleep*, 34:1233–1241.

Mitler, M. M., Gujavarty, K. S., and Browman, C. P. (1982). Maintenance of wakefulness test: a polysomnographic technique for evaluation treatment efficacy in patients with excessive somnolence. *Electroencephalography and Clinical Neurophysiology*, 53(6):658–661, June.

Philip, P., Micoulaud-Franchi, J.-A., Sagaspe, P., De Sevin, E., Olive, J., Bioulac, S., and Sauteraud, A. (2017). Virtual human as a new diagnostic tool, a proof of concept

study in the field of major depressive disorders. *Scientific Reports*, 7(1):426–456.

Poursadeghiyan, M., Mazloumi, A., Nasl Saraji, G., Baneshi, M. M., Khammar, A., and Ebrahimi, M. H. (2018). Using Image Processing in the Proposed Drowsiness Detection System Design. *Iranian Journal of Public Health*, 47(9):1371–1378.

Raake, A. (2002). Does the Content of Speech Influence its Perceived Sound Quality? In *LREC*, pages 1170–1776.

Richard, J.-B., Andler, R., Cogordan, C., Spilka, S., Nguyen-Thanh, V., and groupe Baromètre de Santé publique France 2017. (2019). Alcohol consumption in adults in France in 2017. *Bulletin Epidemiologique Hebdomadaire*, 5-6:89–97.

Sangal, R. (1999). Subjective sleepiness ratings (Epworth sleepiness scale) do not reflect the same parameter of sleepiness as objective sleepiness (maintenance of wakefulness test) in patients with narcolepsy. *Clinical Neurophysiology*, 110(12):2131–2135, December.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., and Krajewski, J. (2011). The INTERSPEECH 2011 Speaker State Challenge. In *Interspeech*, pages 3201–3204.

Schuller, B., Steidl, S., Batliner, A., Schiel, F., Krajewski, J., Weninger, F., and Eyben, F. (2013). Medium-term speaker states-A review on intoxication, sleepiness and the first challenge. *Comput. Speech Lang.*, 28(2):346–374.

Schuller, B., Batliner, A., Bergler, C., Pokorny, F. B., Krajewski, J., Cychocz, M., Vollman, R., Roelen, S.-D., Schnieder, S., Bergelson, E., Cristia, A., Seidl, A., Warlaumont, A., Yankowitz, L., Nöth, E., Amiriparian, S., Hantke, S., and Schmitt, M. (2019). The INTERSPEECH 2019 Computational Paralinguistics Challenge: Styrian Dialects, Continuous Sleepiness, Baby Sounds & Orca Activity. In *Interspeech*.

Schweitzer, J. B., Cummins, T. K., and Kant, C. A. (2001). Attention-deficit/hyperactivity disorder. *The Medical Clinics of North America*, 85(3):757–777.

Sedgwick, P. M. (1998). Disorders of the sleep-wake cycle in adults. *Postgraduate Medical Journal*, 74(869):134–138.

Azmeh Shahid, et al., editors. (2011). *STOP, THAT and One Hundred Other Sleep Scales*. Springer-Verlag New York.

Sparrow, A. R., LaJambe, C. M., and Van Dongen, H. P. (2019). Drowsiness measures for commercial motor vehicle operations. *Accident Analysis & Prevention*, 126:146–159.

Verdot, C., Torres, M., Salanve, B., and Deschamps, V. (2013). Children and adults body mass index in France in 2015. Results of the ESTEBAN study and trends since 2006. *Bulletin Epidemiologique Hebdomadaire*, 13:234–241.

Zigmond, A. S. and Snaith, R. P. (1983). The hospital anxiety and depression scale. *Acta Psychiatrica Scandinavica*, 67(6):361–370.

Åkerstedt, T. and Gillberg, M. (1990). Subjective and objective sleepiness in the active individual. *Int J Neurosci*, 52:29–37.

Åkerstedt, T., Anund, A., Axelsson, J., and Kecklund, G. (2014). Subjective sleepiness is a sensitive indicator of insufficient sleep and impaired waking function. *Journal of sleep research*, 23(3):240–52.