Towards automatic sleepiness measurement through speech

Vincent P. Martin

Towards automatic sleepiness measurement through speech

Vincent P. Martin[1], Jean-Luc Rouas[1], Pierre Thivel[1],
Jean-Arthur Micoulaud Franchi[2], Pierre Philip[2]

[1] Univ. Bordeaux, CNRS, Bordeaux INP, LaBRI, UMR 5800, F-33400, Talence, France.
[2] USR CNRS 3413 SANPSY, CHU Pellegrin, Université de Bordeaux, Bordeaux, France.

Correspondance should be adressed to : vincent.martin@labri.fr

Abstract

Following patients with chronic sleep disorders involves multiple appointments between doctors and patients. Speech technologies and virtual doctors can help reduce the number of appointments. However, there are still some challenges to overcome: Sleepiness measurements are diverse and are not always correlated and most past research focused on raw performance rather than interpretability. We introduce a new database, based on the Multiple Sleepiness Latency Test, that includes several sleepiness measures and a new simple set of features that we test on the Sleepy Language Corpus.

1 Introduction

One of the major challenges for treating neuropsychiatric pathologies is the follow-up of chronic patients in order to adapt treatment and measure early relapses. Such a 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 subjective 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 in response to treatment. The growing number of patients however 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 techniques: looking at eye movements and examining verbal expressions or body movements. Thanks to recent advances in speech processing, it is now possible to detect precise cues in the voice allowing to characterise the state of a speaker (*e.g.* to measure the sleepiness level). This method has multiple advantages as recording voice data is not invasive and it requires neither 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.

Studies on sleepiness detection through voice has seen a peak of interest in the early 2010s, culminating with the 2011 Interspeech challenge [1]. Since then, there have only been few reported research on the subject [2]. If most of these works are based on the Sleepy Large Corpus (SLC) [3, 4], some other propose results on their own database. The comparison of these databases with the SLC mostly suffers from two differences. First, they rely on few subjects: to our knowledge, except from the SLC, only [5] present results over a large database (55 subjects). Second, the evaluation of the sleepiness state is done either by the Karolinska Sleepiness Scale (KSS) [6, p. 209] as in the SLC, the Standford Sleepiness Scale [6, p. 369] as in [5, 7], the Karolinska Drowsiness Test [8] or Electroencephalography (EEG) [9, 10, 11, 12].

In [13], it is shown that using a virtual doctor as a diagnostic tool using questionnaires is well accepted by the patient. We wish to complete the analysis carried out using this method by recording the voice of the patients to determine their sleepiness level. However, there is a few challenges to overcome to successfully reach that goal: there are subjective and objective methods to measure excessive daytime sleepiness and they do not necessarily measure the same symptoms [6]. Moreover, determining which vocal features should be used according to each kind of sleepiness measurement is a quite complex task, especially given

the fact that we wish to use features that can be interpretable by clinicians if further enquiry is needed. Finally, fatigue, depression and sleepiness can be clinically misclassified unless a physician makes a clear investigation of the three symptoms which is rarely the case.

In this paper, we provide leads to reach all the targets mentioned above. First, we introduce a new large database, based on the Multiple Sleepiness Latency Test (MSLT) [14, 15]. The database provides sleepiness measurements at different time granularity: from long term, using self-reported questionnaires on the sleepiness habits of the patients, to instantaneous measures such as the MSLT iteration measure or the KSS [6, p. 209]. Moreover, we propose at the same time objective (iteration value) and subjective (KSS and cartoon faces [6, p. 277]) instantaneous measurements of sleepiness for highly phenotyped patients, allowing to unravel the influence of the individual factors on voice. This database is however still under construction, with a growing number of recordings, but not yet enough for a throughout analysis of all factors and sleepiness measures.

This is why our baseline system is currently built using the SLC database which has both the advantages of having many speakers and being widely used. Working with this database, we propose new simple features to estimate the sleepiness state through voice. Most researches, including the best state-of-the-art system [16], use features computed with openSmile [17]. While some have tried to suggest using different features [4, 18], the common drawback to all these systems is the lack of possible interpretation of the features, due to the use of feature combination techniques. In this paper, we show that by carefully selecting a small set of features that can be interpreted by doctors, we can design a system that achieves classification performances on par with the state of the art.

Although we previously mentioned that our database is not yet complete, we also propose preliminary results using the same systems built on the SLC and discuss them.

The paper is structured as follows. In Section 2, we provide a description of the new introduced MSLT database. In Section 3, we quickly present the SLC corpus and the sub-corpus used in this paper as well as the new features. Section 4 describes the systems elaborated on the SLC, their results, and proposes a brief physiological interpretation of the selected features. Section 5 deals with the results on the MSLT database. Finally, conclusions and future work are presented in Section 7.

2 MSLT database

This section introduces the MSLT database, composed of on-going recordings of speech samples collected at the Bordeaux University Hospital Sleep Clinic, France from patients having a suspicion of excessive daytime sleepiness related to hypersomnia or nocturnal breathing disorders. The statistics of the current state of the database are presented in Table 1.

2.1 Recording Set Up

The MSLT is a sleep recording conceived to diagnose excessive daytime sleepiness and hypersomnias such as narcolepsy. Concretely, the patients are welcomed the evening prior to the exam to be recorded with a first night of polysomnography (PSG) [19]. The day of the exam, they are asked to take a nap every two hours at 9am, 11am, 1pm, 3pm and 5pm. Before each nap, approximately 10 min before the beginning of the exam, we record speech, and the patient fills the KSS questionnaire with the procedure below. After the medical assistant has asked the patient about their level on the Cartoon Sleepiness Scale, the test begins. After switching the lights off, one epoch of stage 1 or one epoch of any other sleep stage is required to establish sleep onset [14]. The maximal value of the MSLT nap is 20 minutes.

Each patient reads six different texts, that are the same at constant session. A reference recording is done the day before at 6pm, time at which patients are not supposed to be sleepy [20]. The procedure is the following. First, the patient is asked to read the text. Second, the patient fills a KSS questionnaire. Third, the patient is asked to read the text aloud and its voice is recorded.

The patients are installed either at their desk or in their bed, the position and orientation of the microphone being the same for all the iterations. The microphone is an omnidirectional Audio-technica AT4022, connected to a audio recorder Tascam DR-100 MKIII.

2.2 Choice of the texts

We choose to use *Le Petit Prince* of Antoine de Saint-Exupéry as it is a children book, it is written in simple grammar and it uses common words. This text does

not convey much emotions, but still, it is not boring. The patients reaction to the texts are either being interested when recognising the origin of the text or no peculiar interest at all. To ensure that we have enough vocal content, we choose to segment the text into sections of approximately 200 words, *i.e.* audio file between 50s and 2min depending on the reading level of the reader.

To take into account the reading proficiency of the patient, we measure their fluency during each reading session. The ELFE measure (Évaluation de Lecture en FluencE) is defined by the number of words read during one second minored by the number of errors [21]. Self-corrected errors are not taken into account as the time spent to correct the errors already penalises the score.

2.3 Patients profiles

The database benefits from two main advantages unseen in previous sleepy speech databases. First, the patients are as phenotyped as possible. Multiple factors possibly affecting voice are measured for each patient: age, sex, height, weight, body mass index, neck size, number of cigarettes and alcohol glasses per day, education level. Medical questionnaires on the behavorial sleepiness and life habits of the patients are also taken into account. A PSG is conducted the night before the exam, fatigue, snoring, hypertension, SOAS, ESS[6, p. 149], ISI[6, p. 191], HAD[22], FOSQ-10[6, p. 179], part A of ASRS[23], FSS[6, p. 167], CAGE[6, p. 415], CDS5 [24], Toronto[6, p. 391], Barcelona[25], Hobson scale [26] are all measured. To complete this profile, we also take into account pathologies and treatments that can affect the voice (psychostimulants, myorelaxants, ...). By working with highly phenotyped patients, this design can account for several biases that may affect speech apart from than sleepiness. However, the drawback is that to analyse all the possible factors, we need to record many patients for the results to be statistically viable.

Instead of only focusing on subjective or objective sleepiness measures, the KSS, the cartoon sleepiness scale and the MSLT iteration values are collected. The KSS value of the SLC is the mixture of subjective and external measure of the sleepiness (see Section 3.3): even if other hybrid metrics such as the ODSI [27] show relevance for sleepiness measure, our design has the advantage to evaluate the influence of each component.

Sex	**Male**	**Female**	**Total**
#subjects	20	31	51
#samples	94	142	236
mean Age (std)	37.70 (16.63)	36.03 (11.80)	36.03 (11.80)
mean KSS (std) (302 samples)	4.03 (1.77)	4.03 (1.77)	4.40 (1.99)
mean MSLT (std) (236 samples)	8.99 (6.25)	12.45 (6.28)	11.07 (6.49)
mean Cartoon faces (std) (251 samples)	1.52 (0.89)	1.72 (0.88)	1.64 (0.89)
MSLT nap ≤ 8	52	45	97
MSLT nap > 8	25	58	83
MSLT nap $= 20$	17	39	56

Table 1: Current state of the MSLT database as of May 2019: mean value followed by standard deviation

3 Baseline systems

Given the fact that the MSLT database does not yet contains a sufficient number of speakers, and in order to easily compare our results with state-of-the-art systems, we decided to build our baseline systems using the SLC.

3.1 SLC Database presentation

Used as reference corpus for the Interspeech 2011 challenge [1] on detection of sleepiness through voice, the SLC is a database composed of multiple speech tasks conducted in parallel of other sleep-deprivation studies by 99 speakers. They are German volunteers and all the speech samples are in German and English. All the details about the dataset and the experimentation can be found in [9, 28, 29].

The other information given in the database are the task, genre, attribution in train-development-test set and the mean of three KSS scores, explained in Section 3.3.

3.2 Selection of a read subset of the database

As most of the patients that are recorded in the MSLT corpus already have sleepiness complaints, and to ensure valid comparison between the two corpuses, we decided to focus on reading tasks. As shown in [30], reading tasks have a lighter cognitive load than spontaneous speech tasks, the latter having the inconvenient of being biased by emotions.

To study the minimum duration required to have enough information to evaluate the sleepiness state of the speaker, we select the reading tasks of the

SLC with a mean duration higher than 8s: *northwind*, *flight1&2* and *roger1*. Then, we extract X_i, the set of features computed on the i first seconds of the file, and X_{i+1}. The features are scaled and the cosine similarity between X_i and X_{i+1} is computed. The cosine similarity measures the similarity between two features. When the cosine similarity reaches 1 and becomes stationary, the length of the audio file is considered long enough to ensure that the features do not evolve. The Figure 1 shows the Mean and Standard Deviation of the evolution of the features when repeating this step on all the audio files. As the features seem not to evolve a lot for audio files longer than approximately 8s, we comfort our choice to work on the selected tasks of the SLC.

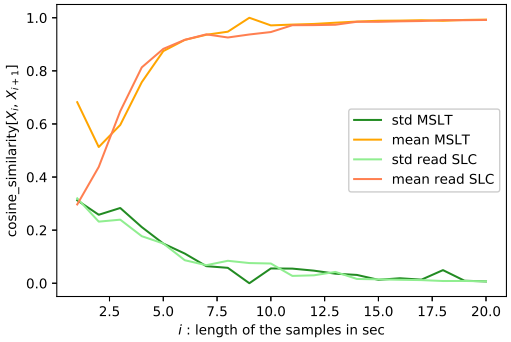

Figure 1: Evolution of the features in function of the duration of the sample, for reading tasks of the SLC and the MSLT database. Mean and Standard Deviation are computed on the different samples of length i. Both the mean and the std suggest that the features do not evolve for duration greater than 15-20s.

3.3 Ground truth: the KSS

The only sleepiness measure provided in the SLC is the mean between three values of KSS (which are instantaneous measurements): one is filled by the patient himself, and two are filled by external trained annotators. The dataset is then split into two classes: following [16, 18, 31, 32], samples with a KSS>7.5 are considered as Sleepy Language (SL) while samples with a KSS≤7.5 are labelled as Non Sleepy Language (NSL). Some statistics on the SLC database with the KSS limit fixed at 7.5 are presented in Table 2. Others choices for this boundary may be made, such as a limit of 7 as chosen in [33] or 8 in [28]. However, setting the limit between sleepy and non-sleepy language is a choice that may be different according to sleepiness specialists. The effect of using different limits is discussed in Section 4.4.

		Train	Dev	Test	All
Male	NSL	58 (1089s)	36 (680.5s)	60 (1035.3s)	154 (1872.8s)
	SL	29 (364.9s)	47 (680.5s)	22 (297.0s)	98 (2804.8s)
Female	NSL	119 (1624.6s)	101 (1286.6s)	93 (1340.7s)	313 (1342.4s)
	SL	96 (941.7s)	63 (753.7s)	66 (806.9s)	225 (4251.9s)
Total		303 (4019.6s)	247 (2720.8s)	241 (3478.9s)	79110 (219.3s)

Table 2: Number of samples and total duration of the selected read parts of the SLC database with KSS limit fixed at 7.5 (59 readers).

3.4 Features extraction

3.4.1 Custom Set of Features

The goal of the project is to find vocal biomarkers that can be understood and used by physicians. As a consequence, several features are extracted. On one hand, some features are computed directly on each recording using either an automatic vocalic segments detection algorithm [34] or voiced segments detected using a fundamental frequency extraction algorithm [35]. These features are: the duration of voiced parts, the percentage in duration of voiced parts, the duration of vocalic segments, the percentage in duration of vocalic segments.

On the other hand, other features are computed on each voiced segment to characterize the regularity of production of harmonic sounds. These features are averaged for each recording. An exhaustive description of these custom features can be found in [36, 37, 38]. They are divided in three classes: measurements (mean, var, max, min, extend) on the fundamental frequency and intensity; descriptive values (frequency, power, bandwidth) of harmonics and formants; cepstral peak prominence (CPP) and HNR computed using a modified version of the Covarep matlab toolkit [39]. The total number of features in the Custom Set of Features (CSF) is 44.

3.4.2 OpenSMILE IS11 features

For comparison purpose, we also extract the most widely adopted feature set consisting of 59 low-level descriptors (4 energy related descriptors, 50 spectral descriptors and 5 voice related descriptors), combined with 33 base functionals and 5 F0 functionals, leading to a total of 4368 features. A complete description of these features is presented in [17].

4 Proposed systems

Hereafter, we propose to test different systems that aim to classify automatically Sleepy Language and Non Sleepy Language on the SLC.

4.1 Feature selection by statistical methods and SVM

Contrary to classical dimension reduction techniques (such PCA or LDA), our framework constrains the selection of features to keep the explicit meaning of the vocal features, so that they can be interpreted by physicians. As a consequence, we choose the features that have the highest correlation to the KSS measures and give good classification results.

Before further analysis with SVM, all the features are centered with the mean values of the features for all the samples of a given speaker (number of speech samples for each speaker: Mean=13.4, StD=19.4) and scaled. After a Shapiro test that ensures that the data is not normally distributed, we conduct a Spearman ρ test to measure the correlation between each of the features and the KSS values. This computation is done only on the aggregated *train+dev*.

Then, we perform a grid search on the two parameters of the system: the KSS limit and the number of features. For each number n of features, we keep the n features that correlate the most with the KSS. The train set is then splitted according to the various possible KSS limits (between 5: "neither sleepy neither awaken" and 9: "very sleepy with great efforts to stay awake"). A SVM classifier with the Radial Basis Function (RBF) kernel is implemented with the Python library *sklearn* [40]. It is trained using the training set and tested against the development set. Finally, for a given set of n features, the reported result will be the best result over all the KSS limit values. This procedure is carried out using our complete set of 44 features and the 100 openSmile features that correlate the most with KSS. On the development set, the best performances are obtained using 23 features from our Custom Set of Features (68.1% of UAR) and 59 features from IS11 openSMILE set (65.0% of UAR) respectively referred as (a) and (b).

The 23 selected features from our set are presented in Table 3 with their Spearman correlation value, their significance and the order in which they are added to the system. An interpretation of the observed trend is discussed in Section 4.5.

Features	**Spearman ρ**	**p-value**	**rank**
durvoiced	0.057	0.17	23
durvowel	0.045	0.29	19
F0 Mean (vowels)	-0.32	***	1
F0 Mean	-0.27	***	2
F0 Slope	-0.085	*	16
F0 Min	-0.20	***	4
F0 Max	-0.24	***	3
F0 Extend	-0.08	0.053	17
Energy Var (vowels)	-0.07	0.1	21
Energy Var	-0.07	0.1	22
Energy Slope	0.13	**	10
Energy Min	0.14	***	8
Energy Extend	-0.16	***	6
H1	0.10	*	14
H2	0.13	**	9
A2	-0.076	0.073	18
A3	-0.10	*	13
F1	-0.19	***	5
B1	-0.10	*	15
H1A1	0.073	*	20
H1A2	0.12	**	11
H1A3	0.14	***	7
HNR05	-0.12	**	12

Table 3: The selected 23 features achieving the best performances, their Spearman ρ, p-value and rank. p<0.05 :*, p<0.01 :**, p< 0.001 :***

4.2 ASIMPLS

Based on the Partial Least Square (PLS) algorithm [41], the ASIMPLS classifier achieve the best results of the state of the art when fused with SVM [16]. The latter article discusses, assuming few hypothesis, that the ASIMPLS has the advantage of taking the different speaking styles into account: speaker normalisation is not needed in this framework.

The two parameters of the ASIMPLS classifier are the number of components k and the correction term b'. The algorithm is the following: First, the features are scaled and normalised. Second, k and b' are tuned on the *train* vs *dev* paradigm and choose the parameters achieving the best accuracy with a KSS limit of 7.5. The result of the selection of k is presented in Figure 2. The two best systems are observed for IS11 openSMILE features, on the whole SLC (k = 5, $b' = 92.8 \times 10{-}3$, UAR = 66.4%) and the reading tasks (k = 10, $b' = 64.7 \times 10^{-3}$, UAR = 67.0%), respectively noted f) and d). We also keep the system using our Custom Set of Features that achieves an UAR of 58.9% with k=5 components and $b' = 4.5 \times 10^{-3}$, noted c).

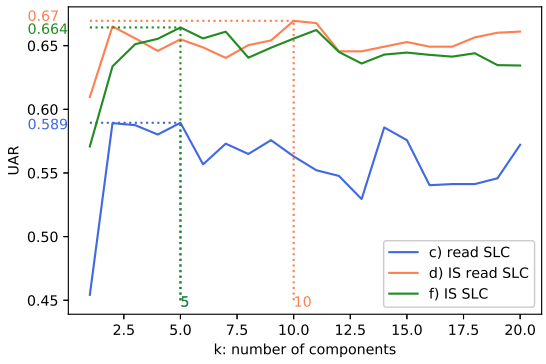

Figure 2: Performances of the ASIMPLS systems with the best b' depending on the number of components on *train* vs *dev*

4.3 Results using the conventional KSS limit

The results of the different systems on *train+dev* vs *test* for a KSS limit of 7.5 are aggregated in the Table 4. On the reading tasks, the best system is the SVM with the custom set of features a). However, the system b) achieves better results than the state of the art. Focusing on the reading tasks, and therefore on longer samples, allows the features to converge in accordance with the results of Section 3.2. On the contrary, the ASIMPLS algorithm achieves better performances on the whole SLC as it is assumed that the multiplicity of samples allow the algorithm to model with better precision the speakers vocal characteristics.

Ref	Features (#)	Sensibility	Specificity	UAR
Reading SLC (241 samples)				
a)	SVM CSF (23)	75%	77.78%	**76.39%**
b)	SVM IS11 (59)	43.1%	86.9%	65.0%
c)	ASIMPLS CSF (5)	44.3%	83.0%	63.7%
d)	ASIMPLS IS11 (10)	56.8%	78.4%	67.6%
Entire SLC database (2808 samples)				
e)	SVM CSF (23)	23.4%	93.1 %	58.3%
f)	ASIMPLS IS11 (5)	67.8%	68.8%	68.3%
	State of the art [16]	64.3%	79.1%	71.7%

Table 4: Results of the systems on *train+dev* vs *test* for KSSlimit=7.5.

4.4 Sensibility of the systems to the KSS limit

To our knowledge, the sensibility of the systems to the KSS limit has only scarcely been studied. Are the designed systems specific to a precise sleepiness state or does it detect more general sleepiness? To answer this question, we study the influence of the KSS limit on the performances of the system. Each previously selected system is trained and tested with SL and NSL differentiated by the KSS limit in the interval [5,8.5]. The results are drawn on Figure 3. The SVM systems are specialised to a range of sleepiness while the ASIMPLS systems have lower performances but are less specific.

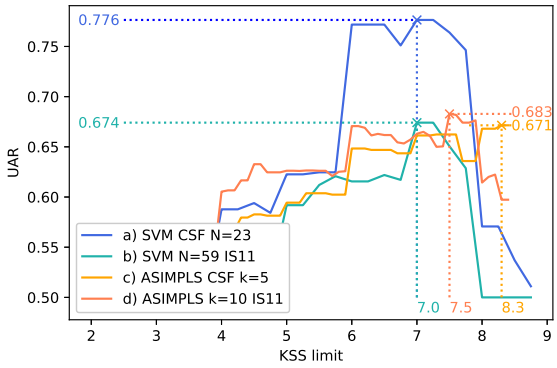

Figure 3: Sensibility of the systems to the KSS limit

4.5 Physiological interpretation of the selected features

One of the biggest constraints of this work is to select features that can relate the physiological modifications of the patient voice caused by sleepiness.

Similarly to [10], an augmentation of the voiced and vowels parts is observed. This observation can be clue to the augmentation of hesitations of the sleepy speakers. The diminution of the values of F0 Mean, F0 Min, F0 Max, F0 Extend, F0 Slope, frequency F1 (also observed in [28, 42]), bandwidth of F1 and amplitude of second and third formants witness a shift of the frequencies contained in the voice towards lower values as already observed in [5, 28, 43]. Moreover, the diminution of the values of F0 Extend and F0 Slope are clues of a reduction of the bandwidth used during the vocal process.

Contrary to the observations made in [28], the energy extend, the absolute value of the energy slope and the variance of the energy decrease with sleepiness. Added to the rise of the low frequencies, and with the high energies staying constant, these observations express a diminution of nuances in the Sleepy Language. We hypothesise that the slight augmentation of the first harmonic frequency, that seems contrary to previous observations, is due to the modification of the exhaled air flux that modifies the distribution of harmonics but not formants [44]. This is consistent with the diminution of the HNR (HNR05 in our case) also observed in [11].

These observations lead to the hypothesis that the sleepy speakers struggle to produce the same variety of nuances of frequencies, energy and quality of voiced parts.

5 First results on the MSLT database

This Section introduces the results of the previously presented systems on the MSLT database. The following results are discussed in the next Section.

5.1 Proposed system on the SLC

In the same fashion than the best systems of the previous parts, we set the KSS limit to 7. The MSLT limit is set to 8 as it is a pathological threshold for excessive daytime sleepiness [45]. To mimic the SLC KSS score which is a mixture of subjective and objective values, a PCA is trained with the MSLT and KSS value, creating a mixed score. Its limit is set to the PCA projection of (MSLT lim=8, KSS lim=7) = 3.05.

First, we train a SVM with the same 23 features computed on the whole SLC on (a) and test it on the whole MLST database (*train+dev+test* SLC vs *train+test* MSLT). This leads to an UAR of 55.2% for the KSS, 48.3% for the MSLT and 50.7% for the PCA score. As both systems are about reading tasks with sufficient length, the results in the same order as on the SLC were expected on the MSLT database. The KSS gives better accuracies than the MSLT measure or the PCA score, which is consistent with the fact that the SLC is trained with its KSS. To test the hypothesis that the differences come from the tuning of the parameters, we decide to apply the same systems as on the SLC but with parameters tuned on the MSLT.

To that end, we choose the reported best systems on the reading tasks with our CSF. On one hand, we apply almost the same procedure as (a): without speaker normalisation, we tune the number of optimal number of features on a SVM (rbf Kernel, C=10) with the same 23 previously chosen features, computed on the MSLT database. This procedure is done under Leave One Speaker Out Cross Validation (LOSOCV) to avoid over-fitting. The confusion matrix is computed for each iteration and the UAR is computed on the summed matrix at the end of the cross validation. The best results are achieved for 6 features: the KSS score leads to an UAR of 65.0%, 67.8% for the MSLT score and 67.6% for the PCA score.

On the other hand, we train and test the ASIMPLS on the MSLT database (*train* vs *test* under LOSOCV to fit the parameters, train vs test under LOSOCV to evaluate the system). This leads to 47.9% UAC for the KSS limit, 60.2% for the MSLT and 58.5% for the PCA score. These performances shows that the differences are partly due to the tuning of the parameters but also come from a lack of similarity between the two databases that is discussed in the following section.

5.2 Discussion

The poor performances of transferring systems tuned on the SLC to the MSLT database show differences between the data. We assume they are due to the differences in the experimental set up between the two databases. The first difference appears in the efforts made to isolate sleepiness from other factors in the MSLT database. While the SLC relies on the maintenance of wakefulness and high cognitive load tasks, the MSLT test is based on the incitement to sleep. Vocal features recorded in these conditions may be linked to sleepiness but also other factors such as fatigue. Moreover, as emotion such as stress has an important influence on the vocal features [46], the patients recorded for the MSLT are welcomed the day before the exam to have time to be acclimated to the place: the vocal content of the SLC is linked to sleepiness but is also polluted by the inherent stress of being recorded in an unfamiliar environment.

Second, the measures of sleepiness are different. Although we attempted to combine the MSLT score and the KSS of the MSLT database, the ground truth on sleepiness state has to be discussed: are the vocal features linked to objective or subjective sleepiness? We have proposed a PCA to elaborate a mixed score

but a deeper study should bring to light the specificity of vocal features to the subjective or objective sleepiness scales. A larger set of data on the MSLT integrating inter-individual differences could refine the results.

Finally, the speaker normalisation seems to be relevant concerning the SLC but not the MSLT database: in the first one, the recordings are temporally close and the state of the speaker doesn't vary from one recording to another, while in the second one the state of the patients can vary a lot (Mean of StD of MSLT: 4.0).

6 Conclusion & Perspectives

In this paper, we have proposed a novel strategy for sleepiness detection in voice, with possible applications in the medical field. We have shown that selecting only reading tasks with sufficient length (sup8s) leads to better results for sleepiness detection.We also proposed a set of features that can be interpreted by physicians. Our system performances are comparable to state of the art methods. The careful selection of features as well as the choice of a subset of the SLC enhance the detection of sleepiness through voice. Moreover, we have proposed a physiological analysis of the vocal parameters for various levels of sleepiness.

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. Further works on this database will include the use of all the medical information to take into account their influence on the voice; a deeper study on the creation of relevant score for the sleepiness detection in voice; the study of the database through a longitudinal approach.

7 Acknowledgements

This work is carried out in the framework of the ISOSA project funded by the French Region Nouvelle Aquitaine and by the SOMVOICE project sponsored by the Labex BRAIN. The authors thank Pr. Jarek Krajewski, Engineering Psychology - Rhenish University of Applied Science (Cologne, Germany), for having provided the SLC database.

References

[1] B. Schuller, S. Steidl, A. Batlinder *et al.*, "The INTERSPEECH 2011 Speaker State Challenge," in *Interspeech*, 2011.

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

[3] D.-Y. Huang, Y. Tsao, H. Chiori *et al.*, "Feature Normalization and Selection for Robust Speaker State Recognition," in *IEEE - International Conference on Speech Database and Assessments*, 2011.

[4] J. Krajewski, S. Schnieder, D. Sommer *et al.*, "Applying multiple classifiers and non-linear dynamics features for detecting sleepiness from speech," *Neurocomputing*, vol. 84, pp. 65–75, 2011.

[5] E. L. McGlinchey, L. S. Talbot, K.-h. Chang *et al.*, "The Effect of Sleep Deprivation on Vocal Expression of Emotion in Adolescents and Adults," *Sleep*, vol. 34, pp. 1233–1241, 2011.

[6] A. Shahid, K. Wilkinson, S. Marcu *et al.*, *STOP, THAT and One Hundred Other Sleep Scales*, 2011.

[7] J. Krajewski and B. Kroger, "Using prosodic and spectral characteristics for sleepiness detection," in *Interspeech*, 2007.

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

[9] M. Golz, D. Sommer, M. Chen *et al.*, "Feature Fusion for the Detection of Microsleep Events," *Journal of VLSI Signal Processing*, vol. 49, pp. 329–342, 2007.

[10] L. S. Dhupati, S. Kar, A. Rajaguru *et al.*, "A novel drowiness detection scheme based on speech analysis with validation using simultaneous EEG recordings," in *IEEE - Int. CASE*, 2010, pp. 917–921.

[11] S. Boyer, R. El-Yagoubi, M. Tiberge *et al.*, "Paramètres Acoustiques de la Voix et Privation de Sommeil," in *CFA/VISHNO*, 2016.

[12] A. R. Sparrow, C. M. LaJambe, and H. P. Van Dongen, "Drowsiness measures for commercial motor vehicle operations," *Elsevier*, 2018.

[13] P. Philip, J.-A. Micoulaud-Franchi, P. Sagaspe *et al.*, "Virtual human as a new diagnostic tool, a proof of concept study in the field of major depressive disorders," *Scientific Reports*, vol. 7, no. 1, pp. 426–456, 2017.

[14] M. R. Littner, C. Kushida, M. Wise *et al.*, "Practice Parameters for Clinical Use of the Multiple Sleep Latency Test and the Maintenance of Wakefulness Test," *Sleep*, vol. 28, no. 1, pp. 113–121, 2005.

[15] D. L. Arand, M. H. Bonnet, T. Hurwitz *et al.*, "The clinical use of the MSLT and MWT," *Sleep*, vol. 28, no. 1, pp. 123–144, 2005.

[16] D.-Y. Huang, Z. Zhang, and S. S. Ge, "Speaker State Classification Based on Fusion of Asymmetric Simple Partial Least Squares (SIMPLS) and Support Vector Machines," *Comput. Speech Lang.*, vol. 28, no. 2, pp. 392–419, 2014.

[17] F. Eyben and B. Schuller, "Opensmile," *ACM SIGMultimedia Records*, vol. 6, pp. 4–13, 2015.

[18] C. Sezgin, B. Günsel, and J. Krajewski, "Medium term speaker state detection by perceptually masked spectral features," *Speech Communication*, vol. 67, pp. 26–41, 2015.

[19] Medical Advisory Secretariat, "Polysomnography in patients with obstructive sleep apnea: an evidence-based analysis," *Ontario Health Technology Assessment Series*, vol. 6, no. 13, pp. 1–38, 2006.

[20] P. M. Sedgwick, "Disorders of the sleep-wake cycle in adults," *Postgraduate Medical Journal*, vol. 74, no. 869, pp. 134–138, 1998.

[21] Cogniscience, "E.L.FE - Évaluation de la Lecture en FluencE," Tech. Rep., 2008.

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

[23] J. B. Schweitzer, T. K. Cummins, and C. A. Kant, "Attention-deficit/hyperactivity disorder," *The Medical Clinics of North America*, vol. 85, no. 3, pp. 757–777, 2001.

[24] D. Courvoisier and J.-F. Etter, "Using item response theory to study the convergent and discriminant validity of three questionnaires measuring cigarette dependence." *Psychology of Addictive Behaviors*, vol. 22, no. 3, pp. 391–401, 2008.

[25] M. Guaita, M. Salamero, I. Vilaseca *et al.*, "The Barcelona Sleepiness Index: A New Instrument to Assess Excessive Daytime Sleepiness in Sleep Disordered Breathing," *Journal of clinical sleep medicine*, vol. 11, no. 11, pp. 1289–1298, 2015.

[26] D. E. Hobson, A. E. Lang, W. R. W. Martin *et al.*, "Excessive daytime sleepiness and sudden-onset sleep in Parkinson disease: a survey by the Canadian Movement Disorders Group," *JAMA*, vol. 287, no. 4, pp. 455–463, 2002.

[27] F. Onen, C. Lalanne, V. M. Pak *et al.*, "A Three-Item Instrument for Measuring Daytime Sleepiness: The Observation and Interview Based Diurnal Sleepiness Inventory (ODSI)," *J Clin Sleep Med.*, vol. 12, no. 4, pp. 505–512, 2016.

[28] J. Krajewski, A. Batlinder, and M. Golz, "Acoustic sleepiness detection: Framework and validation of a speech-adapted pattern recognition approach," *Behavior Research Methods*, vol. 41, no. 3, pp. 795–804, 2009.

[29] B. Schuller, S. Steidl, A. Batlinder *et al.*, "Medium-term speaker states-A review on intoxication, sleepiness and the first challenge," *Comput. Speech Lang.*, 2013.

[30] G. Christodoulides, "Effects of Cognitive Load on Speech Production and Perception," Ph.D. dissertation, 2016.

[31] B. Günsel, C. Sezgin, and J. Krajewski, "Sleepiness detection from speech by perceptual features," in *IEEE - ICASSP*, 2013, pp. 788–792.

[32] Y. Zhang, F. Weninger, and B. Schuller, "Cross-Domain Classification of Drowsiness in Speech: The Case of Alcohol Intoxication and Sleep Deprivation," in *Interspeech*, 2017.

[33] H. Martensson and O. Keelan, "Feature Engineering and Machine Learning for Driver Sleepiness Detection," Ph.D. dissertation, 2017.

[34] F. Pellegrino and R. Andre-Obrecht, "Automatic language identification: an alternative approach to phonetic modelling," *Signal Processing*, vol. 80, no. 7, pp. 1231–1244, 2000.

[35] K. Sjölander, "The Snack Sound Toolkit," 2004. [Online]. Available: http://www.speech.kth.se/snack/

[36] J.-L. Rouas and L. Ioannidis, "Automatic Classification of Phonation Modes in Singing Voice: Towards Singing Style Characterisation and Application to Ethnomusicological Recordings," in *Interspeech*, 2016, pp. 150–154.

[37] J.-L. Rouas, T. Shochi, M. Guerry *et al.*, "Categorisation of spoken social affects in Japanese: human vs. machine," in *ICPhS*, 2019.

[38] V. P. Martin, J.-L. Rouas, and J. Krajewski, "Sleepiness detection on read speech using simple features," in *Interspeech*, 2019.

[39] G. Degottex, J. Kane, T. Drugman *et al.*, "COVAREP — A collaborative voice analysis repository for speech technologies," in *IEEE - ICASSP*, 2014, pp. 960–964.

[40] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[41] S. De Jong, "SIMPLS: An alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, vol. 18, no. 3, pp. 251–263, 1993.

[42] H. P. Greeley, E. Friets, J. P. Wilson *et al.*, "Detecting Fatigue From Voice Using Speech Recognition," in *IEEE International Symposium on Signal Processing and Information Technology*, 2006, pp. 567–571.

[43] T. L. Nwe, H. Li, and D. Minghui, "Analysis and Detection of Speech under Sleep Deprivation," in *Interspeech*, 2006.

[44] J. Hillenbrand, R. Cleveland, and R. L. Erickson, "Acoustic correlates of breathy vocal quality," *Journal of Speech, Language, and Hearing Research*, vol. 37, no. 4, pp. 769–778, 1994.

[45] M. S. Aldrich, R. D. Chervin, and B. A. Malow, "Value of the multiple sleep latency test (MSLT) for the diagnosis of narcolepsy," *Sleep*, vol. 20, no. 8, pp. 620–629, 1997.

[46] J. H. L. Hansen and S. Patil, "Speech Under Stress: Analysis, Modeling and Recognition," in *Speaker Classification I*, C. Müller, Ed., 2007, vol. 4343, pp. 108–137.